Universidad Carlos III de Madrid

Doctoral Thesis

Recombining observations in cluster analysis: The SAGRA method

Author:
Adolfo Álvarez

Supervisor:
Dr. Daniel Peña

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
in the

Graduate Program in Business Management and Quantitative Methods
Department of Statistics

Getafe, May 2014

Acknowledgements

Una tesis no es más que unas páginas en blanco que se van llenando poco a poco con investigación, teoría, muchas horas de frustración, noches sin dormir, pero por sobre todo con el apoyo de mucha gente que en una fase u otra se aparecen en tu camino para darte la palabra justa que necesitabas para seguir escribiendo estas hojas. A todos ellos, gracias.

Agradezco a mi director de tesis, Daniel Peña, por su completa dedicación y apoyo incondicional durante este largo proceso. Sin su ayuda este trabajo no habría sido posible, y le estaré siempre agradecido por haberme dado la oportunidad de aprender de un gran investigador y mejor persona. Gracias querido Daniel.

Agradezco además a todos los miembros del departamento de Estadística de la Universidad Carlos III de Madrid a lo largo de estos años, quienes con su compañía, sabio consejo, o simplemente con una distendida charla a la hora de comer hicieron de mi permanencia aquí una experiencia inolvidable y que atesoraré por siempre. Gracias.

A Visi y a María, por hacer que mi paso por la UC3M fuera un poco más fácil. Gracias por tanto cariño.

Agradezco al departamento de Ingeniería Industrial de la Universidad de Santiago de Chile, quienes desde mi aceptación en el programa me dieron su apoyo para venir y constantemente se han interesado en mis progresos, en especial a su director, Juan Sepúlveda, y a mi anterior director de tesis, Miguel Alfaro. También al departamento de Estadística de la Pontificia Universidad Católica de Chile, y en especial a los profesores Fernando Quintana y Guido del Pino por haberme dado la oportunidad de trabajar con ellos en una estancia de investigación. A mis compañeros en Analyx, es un placer trabajar con vosotros y esta tesis también es vuestra. En especial a Sascha, Maciej, Ewa, Monika, y todo el "Analyx core team". Gracias, Dziekuje!

A mis compañeros de viaje, Mariana, Santiago, Ale, Andrea, Ester, Júlia, Alba, Maye, Audra, Ana María, Zulma, Juliana, Agata, y en especial a quienes por más tiempo han estado siempre en las buenas y en las malas, Ana Laura, Paula y Argyro. Gracias.

Un agradecimiento especial a mi familia, quienes desde el primer día me han apoyado en la locura de cruzar el océano "por un par de años". A Cristina por haber sido parte importante de este viaje, y a mis amigos en Chile Andrés, Nacho, Roxana, Cristian, Rodrigo, Mariela, Emilio, y tantos otros por estar siempre ahí. Gracias.

Finalmente agradezco a Anna por ser mi compañera, gracias por tu paciencia, al final ha valido la pena. Jamás habría podido acabar esta tesis sin ti.

Esta investigación ha sido financiada por el proyecto de investigación SEJ-2007-64500 de la Dirección General de Investigación Científica y Técnica del Ministerio de Economía y Competitividad del Gobierno de España, gracias a la ayuda de investigación FPI código BES-2008-009290.

Todos los errores y omisiones contenidos en este trabajo son de mi exclusiva responsabilidad.

Resumen

El objetivo de esta tesis es discutir y desarrollar métodos de partición y recombinación de conjuntos de datos para encontrar su estructura subyacente. En base a esta definición, los métodos cubiertos aquí pueden clasificarse como de aprendizaje no supervisado, o de análisis cluster, dado que no se dispone de información previa de pertenencia de los datos a grupo alguno. Además, en cuanto al problema de fijar el número de grupos, nuestras propuestas están basadas en métodos que no necesitan conocer de antemano este parámetro.

La idea original que motiva esta investigación viene de algoritmos de análisis cluster como el SAR, propuesto por Peña, Rodriguez y Tiao (2004), el cual divide la muestra en pequeños grupos altamente homogéneos para luego recombinar las observaciones y formar la configuración definitiva de los datos (ver Capítulo 1). Sin embargo, cuando se desea recombinar grupos en lugar de observaciones se tiene el problema de que los grupos conforman particiones disjuntas, y por tanto dependientes entre si, por lo que no pueden aplicarse contrastes tradicionales de igualdad de media o varianzas para su recombinación. Específicamente, esta tesis doctoral quiere contribuir al problema de recombinar pequeños grupos homogeneos para reconstruir la estructura del conjunto de datos.

La tesis está estructurada como sigue: En el capítulo 1 comenzamos estableciendo el marco del problema bajo los métodos de heterogeneidad de modelos y de análisis cluster, revisando alguna de las principales publicaciones en el área. En la segunda parte del capítulo, revisamos el método SAR propuesto por Peña, Rodriguez y Tiao (2004), remarcando algunas definiciones importantes, ejemplos de aplicación y apuntando potenciales mejoras que serán abordadas más tarde.

En el capítulo 2, resumiremos la teoría de los estadísticos de orden, presentaremos nuevos resultados acerca de la distribución triangular y como esta puede ser usada para aproximar una distribución normal. También, abordaremos la combinación

lineal de estadísticos de orden para desarrollar un método de recombinación. Finalmente, consideramos utilizar medidas de profundidad como una extensión natural de los estadísticos de orden.

El capítulo 3 está dedicado a presentar un método univariante para combinar grupos basado en la detección de modas. Comenzamos el capítulo con una breve revisión bibliográfica para luego proponer una metodología de recombinación utilizando el "test dip" elaborado por Hartigan y Hartigan (1985). Posteriormente, se discuten y proponen alternativas para aplicar esta recombinación a estructuras de datos multivariantes.

En el capítulo 4 enfrentaremos el problema de la recombinación en datos multivariantes. Presentaremos un algoritmo basado en el proceso de partición del SAR, mientras que la recombinación se realiza iterativamente utilizando un Factor Bayes, en el que se comparan dos modelos que explican la distribución de los datos dependiendo de las particiones obtenidas.

Para finalizar, en el capítulo 5 resumimos las principales conclusiones de esta tesis, además de presentar algunas líneas abiertas en las cuales basar futuras investigaciones.

Las contribuciones principales de la tesis son las siguientes:

- Un nuevo algoritmo de clustering llamado SAGRA (Splitting And Group Recombining Algorithm), basado en una estrategia de partir y recombinar, usando la función discriminador y un método de deteccion y limpieza de datos atípicos para partir los datos y luego factores de Bayes para recombinar los grupos.

- Formulación de la esperanza exacta y aproximada de estadísticos de orden para la distribución triangular. Estos resultados pueden ser usados para aproximar las esperanzas de estadísticos de orden para una distribución normal.

- Un procedimiento basado en el bootstrap para recombinar particiones univariantes basado en combinaciones lineales de estadísticos de orden.

- Un enfoque basado en medidas de profundidad para recombinar particiones multivariantes.

- Un método para recombinar particiones por pares, usando tests de unimodalidad tanto en datos univariantes como multivariantes, incluyendo una herramienta gráfica para visualizar la evolución de la recombinación.

Contents

List of Figures

List of Tables

A mi madre y a mi hermana.

Goals and structure of the thesis

The main goal of the thesis is to develop methods to split and recombine a data set in order to find its underlying structure. Using this definition, the methods covered here can be classified under cluster analysis, or as data partition methods, since no labels are available for the data. Regarding the number of groups problem, we will base our proposal in methods which do not need to fix previously this parameter.

The original idea for this research thesis came from algorithms for Cluster Analysis like SAR proposed by Peña, Rodriguez, and Tiao (2004), which divide the sample into small highly homogeneous groups and then recombine the observations to form the definitive data configuration (see Chapter 1). The partition process leads to disjoint groups that cannot be recombined using traditional homogeneity of means or variances tests because the assumption of independence does not hold. Specifically, this thesis wants to answer the question of how we can recombine this small homogeneous groups to reconstruct the structure of the data set.

To answer this question, the thesis is structured as follows: In Chapter 1 we will start setting the framework of the problem and reviewing some of the main literature in the area. In a second part of the chapter, we will revisit the SAR Method proposed by Peña et al. (2004), including some application and examples, showing some potential improvements we will address later.

In Chapter 2, we will review the theory of order statistics, that is useful for our purposes. We give some new results about the Triangular distribution and how can be used to approximate a Normal distribution. Later, we will use linear

combination of order statistics to develop a recombination method. Finally, we consider to use depth measures as a natural extension of order statistics.

Chapter 3 introduces an univariate method to merge groups based on unimodality detection. We start the chapter with a brief literature review, to later propose a recombining methodology using the "dip test" elaborated by Hartigan and Hartigan (1985). Subsequently, we discuss and propose alternatives to implement the recombining into multivariate data structures.

In Chapter 4 we will introduce SAGRA (Split And Group Recombining Algorithm), a cluster analysis methodology based on the splitting from the SAR and a new recombining technique using Bayes factors.

Chapter 5 summarizes the main conclusions of the research work, some open lines and problems which can be subject to further research.

Chapter 1

Introduction

1.1 Model heterogeneity

In statistical analysis we refer as a "model heterogeneity problem" when not all the data points in the sample can be explained by the same model. For example, one of the applications of model heterogeneity is the problem of outliers, where most of the data points come from the same distribution but a few of them have been generated by one or several distributions which differ from the previous one.

The existence of model heterogeneity can bring significant complications when performing inference, because biased estimates of the parameters can be obtained, with the consequent loss of efficiency in estimation and a bad prediction.

In multivariate analysis, model heterogeneity has been studied mainly under the name of "cluster analysis", which has as a main objective to arrange the observations into homogeneous groups by defining similarities between them. Commonly, cluster analysis is used to join data points but also it is possible to apply it to arrange variables, as we will see in the next paragraphs.

These methods are also known as "automatic unsupervised classification methods" or "unsupervised pattern recognition methods". The name of "unsupervised" is

used to distinguish them from discriminant analysis, where the researcher known in advance the possible distributions which can generate the data to be classified.

In general, Cluster Analysis deals with three kinds of problems:

- Partition of the data. When the available data are expected to be heterogeneous and it is required to divide them into a number of clusters so that each element belongs to one and only one of the groups and each item is classified. The most famous partition algorithm is the k-means, by MacQueen (1967), which we will review in deep later.
- Construction of hierarchies. Here the aim is to structure hierarchically the elements of a data set by their similarity. Strictly speaking, these methods do not define groups, but they show the structure of chain association that may exist between the elements. However, the hierarchy obtained, also allows to define a partition of the data into groups, deciding to stop at a certain level of the hierarchy. A classical example of this approach is the Ward Method, developed by Ward Jr. (1963).
- Classification of variables. In presence of many variables, it is interesting to make an initial exploratory study to divide the variables into groups. Such studies may be useful as a guide prior to the application of formal models to reduce dimensionality. A recent approach of this method is given by Raftery and Dean (2006).

Particularly on this thesis, we will center the discussion on the data partition case.

1.2 Cluster analysis for data partition

As we have seen before, the main goal of cluster analysis for data partition is to assign each point of the sample to a certain group. According to Gan, Ma, and Wu (2007), the observations in a group should have the following characteristics:

- Share the same or closely related properties
- Show small mutual distances or dissimilarities
- Have contacts or relations with at least one other object in the group
- Be clearly distinguishable from the complement, i.e., the rest of the objects in the data set.

Many methods have been developed to split the data into homogeneous groups with the previous characteristics, but certainly one of the most popular is k-means. Jain (2010), asserts that k-means was independently discovered in different scientific fields by Steinhaus (1956), Lloyd (1982)(proposed in 1957), Ball and Hall (1965), and MacQueen (1967), being the last one the most known and cited article about the method.

Though developed for over 50 years, k-means is still one of the most widely used algorithms in cluster analysis. A quick search on *Google Scholar* articles search engine finds that only during 2012 more than 24000 published articles and patents were related to k-means. Its ease and simplicity, added to the fact that every statistical analysis software contains an implementation of k-means, certainly contribute to its long-term success.

Briefly, given a sample of observations x_i with $i = 1, .., n$, each of them assigned to a certain group k from a total number of groups K, with $K \leq n$, the k-means algorithm is based on minimization of the Sum of Squared Errors, where the error is defined as the discrepancy between one point x_i and the sample mean x_k of the group k where it has been assigned, in this way:

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in k} \|x_i - \bar{x_k}\|^2$$

To reach the minimization criteria, the k-means algorithm needs to fix the number of groups K. Given that, and following Jain and Murty (1999) the procedure is as follows:

1. Choose K cluster centers randomly
2. Assign each observation to the closest cluster center with the Euclidean distance.
3. Recompute the cluster centers using the current cluster memberships.
4. If the convergence criterion is not met, go to step 2.

Because of the popularity of k-means, besides its generalized use, many modifications in order to improve its behaviour has been proposed in literature. Some of these modifications are K-medians (Kaufman and Rousseeuw 1990), Fuzzy C-Means (Dunn 1973) , Dynamical Clustering Algorithm (Diday 1973), or Trimmed k-means (Cuesta-Albertos, Gordaliza, and Matrán 1997; García-Escudero, Gordaliza, Matrán, and Mayo-Iscar 2008). For a deep overview of these and another methods derived from k-means, go to Jain and Murty (1999), Jain (2010), and García-Escudero, Gordaliza, Matrán, and Mayo-Iscar (2010).

Another very popular approach for clustering are the "Model-Based Clustering" methods. (Fraley and Raftery 1998; Raftery and Dean 2006; Yeung, Fraley, Murua, Raftery, and Ruzzo 2001; Banfield and Raftery 1993; Fraley and Raftery 1999; McLachlan and Peel 2004) On these algorithms, it is assumed that the data are generated by a mixture of probability distributions in which each component represents a different cluster (Gan et al. 2007) in the following way:

Denote again $x_1, x_2, ..., x_n$ as a multivariate sample coming from an unknown probability distribution $p(x)$. Assuming the existence of G groups, each of them is represented by a density $p_g(x)$. In this case $p(x)$ can be represented as a mixture of distributions (usually Gaussian):

$$p(x) = \sum_{g=1}^{G} \pi_g p_g(x; \mu_g, \Sigma_g)$$

where π_g is the probability of the point x_i belongs to group g ($\pi_g \geqslant 0; \sum_{g=1}^{G} \pi_g = 1$), and assuming normality for p_g,

$$p_g(x_i|\mu_g, \Sigma_g) = (2\pi)^{-p/2}|\Sigma_g|^{-1/2} \exp\left\{-\frac{1}{2}(x_i - \mu_g)^T \Sigma_g^{-1}(x_i - \mu_g)\right\}$$

In order to estimate the parameters we need to maximize the likelihood of the mixture given by:

$$L(\mu_1, ..., \mu_G; \Sigma_1, ..., \Sigma_G; \pi_1, ..., \pi_G|x) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g p_g(x; \mu_g, \Sigma_g)$$

Given a fixed G, is possible to estimate π_g, μ_g and Σ_g. Also, the authors propose to restrict Σ_g, by using the decomposition $\Sigma_g = \lambda_g D_g A_g D_g^T$, in such a way the number of estimated parameters is diminished. Those components allow to control the orientation (D_g), the shape (A_g), and the volume (λ_g) of the group g.

The estimation is performed using the EM algorithm as follows:

Assuming there is an unobserved component z_i in the data, so $x_i = (y_i, z_i)$, with $z_i = (z_{i1}, z_{i2}, ..., z_{iG})$ and $z_{ig} = \{1 \; if \; x_i \in group \; g \; ; \; 0 \; o.c.\}$.

The model assumptions are that the distribution of an observation x_i given z_i is $\prod_{g=1}^{G} p_g(y_i|\theta_g)_{ig}^{z}$ and then z_i is multinomially distributed, choosing among a group of G categories with probability $\pi_{1,}\ldots, \pi_G$

The log-likelihood of the "complete" data will be:

$$L(\theta_g; \pi_g; z_{ig}|x) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log\left[\pi_g p_g(y_i|\theta_k)\right]$$

Finally the estimation steps are:

Step E

$$z_{ig} \leftarrow \frac{\hat{\pi}_g p_g(x_i|\hat{\mu}_g, \Sigma_g)}{\sum_{j=1}^{G} \hat{\pi}_j p_j(x_i|\hat{\mu}_j, \hat{\Sigma}_j)}$$

Step M

$$n_k \leftarrow \sum_{i=1}^{n} \hat{z}_{ig}; \hat{\pi}_g \leftarrow \frac{n_k}{n}; \hat{\mu}_g \leftarrow \frac{\sum_{i=1}^{n} \hat{z}_{ig} x_i}{n_k}$$

considering $\hat{\Sigma}_g$ depending on the assumed model.

This is estimated for all "possible" values of G. Understanding "possible" as the set of values of G which have a useful meaning for the researcher, usually between 1 and 9, but depending on the aim of the study. Finally the value which maximizes the BIC, over all found solutions is chosen.

As can be seen, cluster analysis under the context of model heterogeneity, covers a wide variety of problems, which in turn can be approached from several viewpoints. K-means and M-clust are two of the most used and famous algorithms, but everyday new cluster methods appear. A good literature review of cluster analysis can be found in Gan et al. (2007).

Therefore, it must be noticed that many of these algorithms (including k-means and Model-based clustering) have an underlying problem to solve: The detection of the number of groups in the data set. We will embrace this topic in the next section.

1.3 Finding the proper number of groups

Hennig (2010b) discusses the definition of clusters assessing that solutions for any cluster analysis method always depend on what kind of clusters the researcher is looking for. For example, clusters can be defined by gaps, or zones with low density of points between more dense areas; or by "patterns" where a cluster correspond to the shape of certain distribution or geometrical pattern.

Finally, the visualisation of the separation between components and assessment of stability of given clusters can help with the decision about the desired number of clusters, and with how the results are interpreted.

Now we will review some approaches to the problem of finding the number of groups, depending on the definition for the desired clusters. In this regard, Hartigan (1975) define the number of clusters "q" in a p-dimensional statistical population, as the number of connected components from the set {f >c}, where f is the underlying density function on $\mathbb{R}^p$ and c is a given constant. A clear, although more restrictive way to see this definition in practice is thinking in terms of the points that maximize the density function f, or modes. In this way it is possible to assimilate the number of groups by finding the modes of the data set.

Cuevas, Febrero, and Fraiman (2000) based their work on a mode based approach to provide some light on the proper definition of the number of clusters in a dataset. In particular, the authors argue that the problem is to estimate the number of connected components $T(S)$, in a level set $S = S(f; c) = \{f > c\}$ from a sample of random variables $X_1, X_2, ..., X_n$ obtained from f, where f is an unknown function in $\mathbb{R}^p$, and $c > 0$ is a given constant as in the Hartigan's definition.

The authors propose to approximate the set of estimated level $\{\hat{f}_n > c\}$ through a simple estimator whose number of connected components can be evaluated by an also simple algorithm. This estimator of the set is an union of spheres centred in the following way:

$$\hat{S}_n = \bigcup_{i=1}^{k_n} B(X_i, \epsilon_n)$$

where $X_i, i = 1, 2, ..., k_n$ are the sample observations such that $X_i \in \{\hat{f}_n > c\}$ and $B(X_i, \epsilon_n)$ is the ϵ_n closed sphere centred at X_i . Finally, the estimator T_n , will be the number of connected components from $\hat{S}_n$:

$$T_n = T(\hat{S}_n)$$

One approach based on modes is proposed by Peña, Viladomat, and Zamar (2012) who introduce an algorithm called ATTRACTORS, which is a modification of the CLUES procedure Wang, Qiu, and Zamar (2007). The algorithm detect the

number of clusters by using a nearest-neighbours approach: a local median is calculated iteratively for each observation, and this sequence converges to a local mode. In this way, all the local modes detected are called "fixpoints" and each cluster is defined by one fixpoints and its corresponding observations, moreover, Tibshirani, Walther, and Hastie (2001) propose another alternative to estimate the number of clusters K, based on the idea that the dispersion within each cluster decreases monotonically when the number of clusters increases, but after a certain value of K, the decline is markedly flat.

Assuming the data are separated in K groups $C_1, C_2, ..., C_k$, being C_r a set of observations in cluster r, and $n_r = |C_r|$, the cardinality of that set, i.e., the number of observations in the group. Let $D_r = \sum_{i,i' \in C_r} d_{ii'}$ be the sum of distances between each pair of points inside cluster r, and defining the dispersion W_k as:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

So, considering "d" as the square of the Euclidean distance , then W_k is the total sum of squares within clusters. The idea proposed by the authors is to standardize the plot of $\log(W_k)$ and comparing it with their expectation under an appropriate reference distribution to the data.

Finally, the estimation of the optimal number of clusters will be the value of k for which $\log(W_k)$ deviates more from its reference curve. Thus, the estimator Gap is defined as:

$$Gap_n(k) = E_n^*\{\log(w_k)\} - \log(W_k),$$

where E_n^* is the expectation under a sample of size n, obtained from a reference distribution function. The estimator of the optimal number of clusters $\hat{k}$ will be the value which maximize $Gap_n(k)$ after taking into account the sampling distribution.

Calinski and Harabasz (1974) propose the Variance Ratio Criterion, which is more commonly known as the "Calinski criterion" based on the sum of squares between (SS_B) and within (SS_W) the groups:

$$VRC_k = (SS_B/(k-1))/(SS_W/(n-k))$$

Anyway to fix a set of possible number of groups is needed, because to determine the final number of groups, he suggest to select k which one minimize $(VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1})$.

Another popular approach used to estimate the appropriate number of components is the BIC criteria used among others, by the model based clustering algorithm (Fraley and Raftery 1998; Schwarz 1978), as we have seen before. Under this framework, is possible to compare different group settings by maximizing the Bayesian Information Criterion (BIC) :

$$G = arg \max_G (2L(G) - r\log(n))$$

Where, $L(G)$ is the log-likelihood of the best G component model, r is the number of parameters of the model and n is the number of observations.

In summary, any cluster analysis method should be able to split the data set under some heterogeneity measure and provide a way to define the number of groups. In the next chapters we will present our proposals including these aspects for multivariate analysis.

1.4 The SAR procedure

1.4.1 Introduction to the SAR procedure: definition of the heterogeneity measure

Peña et al. (2004) propose a new exploratory approach to address the problem of identifying clusters in particular, and model heterogeneity in general. The method, named by authors as SAR (Split And Recombine), divide the sample into smaller subgroups and then recombine their observations to form the final clusters. Furthermore, this methodology is general enough to encompass also problems of identification of outliers, both in multivariate cluster analysis (Peña et al. 2004) and in regression (Rodriguez 2002).

The SAR procedure is based on the concept of model heterogeneity as follows:

Let M be the model adjusted to a set of n observations $X = (x_1, x_2, ..., x_n)$, where x_i is a vector of dimension p. The procedure is based on defining a measure $H(x, X)$ of heterogeneity between an observation x and the data set X, and iteratively use this measure to cover the following steps: To identify outliers and eventually delete them from the sample; to split the sample into more homogeneous groups and finally, recombine the observations to form the final clusters.

To this end, the authors argue that the natural way to test whether a new observation is homogeneous with respect to the rest of data set is to see whether this element is close to its prediction based on X, and the model M, with p-dimensional vector of parameters θ. Then, assuming that for certain θ, observations X and x are independent, the distribution of the prediction for a new data point x given X is equal to:

$$p(x/X) = \int p(y/\theta)p(\theta/X)d\theta,$$

where $p(x/X)$ is the distribution of the data point x , while $p(\theta/X)$ is the posterior distribution for parameter θ. Thus, if the density of the observed value is small, there is a reason to believe that this value is heterogeneous with respect to the sample X.

However, it is not always easy to obtain these distributions, so the authors propose an alternative by normalizing the predictive density over the modal value $\hat{x}$, which yields to the following measure of heterogeneity:

$$H(x, X) = C_0(x) = -2 \ln \frac{p(x/X)}{p(\hat{x}/X)}.$$

Assuming a set of independent observations coming from an univariate normal distribution $N(\mu, \sigma^2)$, where distribution parameters (μ, σ) have non-informative a priori distribution $p(\mu, \sigma) \propto \sigma^{-1}$, , then $\hat{x} = E(x/X)$ and the measure of heterogeneity is defined as:

$$C_0(x) = n \ln \left\{ 1 + \frac{t^2}{\nu} \right\}$$

where $\nu = n - 1, t^2 = \left(\frac{N}{N+1}\right) \frac{(x-\bar{x})^2}{s^2}$, $\bar{x}$ is the sample mean of the N observations on X, and $s^2 = \nu^{-1} \sum_j (x_j - \bar{x})^2$, is the corresponding sample variance. Finally t^2 has a F distribution with 1 and $n - 1$ degrees of freedom.

The previous measure of heterogeneity for the univariate case can be also used as a base to decide if one of the observations from the sample X is homogeneous with respect to the rest, examining how this measure behaves when the data point is deleted from the sample. In this case, the heterogeneity measure is defined as:

$$c_0(i) = (n - 1) \ln \left\{ 1 + \frac{t^2_{i(i)}}{\nu_0} \right\}$$

where $\nu_0 = n-2, t^2_{i(i)} = \left(\frac{n-1}{n}\right)\frac{(x_i-\bar{x}_{(i)})^2}{s^2_{(i)}}$, $\bar{x}_{(i)}$ is the sample mean of the n-1 remaining observations from $X_{(i)}$, obtained by removing $x_{(i)}$ from the sample X, and $s^2_{(i)} = \nu_0^{-1}\sum_{j\neq i}(x_j - \bar{x}_{(i)})^2$ the corresponding sample variance.

In this case, the authors propose that if there is only one outlier in the sample, an effective procedure to detect it, would be to observe if the maximum value of the measure $c_0(i)$ for every $x_{(i)}$ is greater than some cut-off value to decide whether the observation is an outlier . However, this measure has a limitation if there is any other heterogeneous point with the rest of the sample, especially if both are near outliers, a problem called "masking effect" as referred by Murphy (1951), Bendre and Kale (1985) and Bendre and Kale (1987), among others.

To solve this problem, a new measure of heterogeneity is proposed, this time considering a second point as outlier in the sample as follows:

$$c_1(i/j) = (n-2)\ln\left\{1 + \frac{t^2_{i(ij)}}{\nu_1}\right\}, j \neq i \tag{1.1}$$

where $\nu_1 = n-3, t^2_{i(ij)} = \left(\frac{n-2}{n-1}\right)\frac{(x_i-\bar{x}_{(ij)})^2}{s^2_{(ij)}}$, $\bar{x}_{(ij)}$ is the sample mean of the n-2 remaining observations $X_{(ij)}$ obtained from removing $x_{(i)}$ and $x_{(j)}$ from the sample X, and $s^2_{(ij)} = \nu_1^{-1}\sum_{k\neq i,j}(x_k - \bar{x}_{(ij)})^2$ the corresponding sample variance. Similarly to the previous case, instead of considering the measure proposed in (1.1) for the remaining (n-1) cases, it seems to be more efficient to use the maximum of $c_1(i/j)$ over all points $x_j, j = 1, 2, ...n, j \neq i$ as an heterogeneity measure of x_i in relation to all (n-1) subsets of (n-2) remaining observations in the sample:

$$c_1(i) = \underset{x_j}{max} c_1(i/j)$$

Based on these measures of heterogeneity, the authors argue that the difference $d_1(i) = c_1(i) - c_0(i)$ is able to measure the maximum increase in the heterogeneity of x_i with the rest of the sample when another point is excluded. The cut-off values for that measure have been simulated under normality assumptions.

In the case of the multivariate analysis, the above measures are expressed as follows:

$$C_0(i) = (n) \ln \left(1 + \frac{Q}{\nu}\right) \tag{1.2}$$

where n is the sample size, $\nu = n-m$;$Q = \frac{n}{n-1}(x-\bar{x})'\hat{V}^{-1}(x-\bar{x})$; $\hat{V} = \frac{1}{\nu}E'E, \bar{x} = 1'X/n$ and its equivalent c_0:

$$c_0(i) = (n-1) \ln \left(1 + \frac{Q_{i(i)}}{\nu_0}\right)$$

where n is again the sample size, $\nu_0 = n - m - 1$;

$Q_{i(i)} = \frac{n-1}{n}(x_i - \bar{x}_i)'\hat{V}_{(i)}^{-1}(x_i - \bar{x}_{(i)})$; $\hat{V}(i) = \frac{1}{\nu_0}E'_{(i)}E_{(i)}, \bar{x}_{(i)} = 1'X_{(i)}/(n-1)$ and $X_{(i)}$ is obtained removing x_i from X . Finally c_1 will be:

$$c_1(i) = \underset{x_j}{max} c_1(i/j) = (n-2) \underset{x_j}{max} \left\{ \ln \left(1 + \frac{Q_{i(ij)}}{\nu_1}\right) \right\} \tag{1.3}$$

where $\nu = n-m-2$, and $Q_{i(ij)} = \frac{n-2}{n-1}(x_i - \bar{x}_{(ij)})'\hat{V}_{(ij)}^{-1}(x_i - \bar{x}_{(ij)})$ is the Mahalanobis distance calculated removing the i_{th} and j_{th} elements.

1.4.2 The discriminator function

With the aim of split the original data set into smaller groups, after detecting and potentially eliminating the outliers, the authors define x_l as the discriminator of x_i if the latter observation appears as most discrepant (using the heterogeneity measures) with respect to the rest of the data set when the discriminator is deleted from the sample. The underlying idea is the following: If two observations are identical, they must have the same discriminator, thus, if they are close enough to each other, they should still have the same discriminator.

Formally, x_l is the observation which maximizes $c_1(i/j)$ with respect to the remaining data points $x_j(j = 1, 2, ..., n; j \neq i)$ to produce $c_1(i)$:

$$x_l = arg \underset{x_j}{max} \left\{ -2 \ln \frac{p(x/X_(ij))}{p(\hat{x}/X_(ij))} \right\}.$$

which in the multivariate case, assuming normality, is equivalent to maximize the Equation (1.3), and at the same time, equivalent to:

$$x_l = arg \underset{j}{max} \ (x_i - \bar{x}_{(ij)})' \hat{V}^{-1}_{(ij)} (x_i - \bar{x}_{(ij)}) \tag{1.4}$$

which is the Mahalanobis distance between the element x_j and the rest of the sample, when the i_{th} and j_{th} elements are removed.

In the univariate case, Peña et al. (2004) shows that the discriminator are always the extreme points, while in the multivariate case, Rodriguez (2002) generalize this result demonstrating that the discriminators belong to the convex hull of the sample. Therefore, Rodriguez (2002) proofs that discriminators are invariant to scale and positions transformation, because they are a monotonic function of the Mahalanobis Distance. Using these two properties, the observation x_l will be the discriminator of x_i if and only if:

$$x_l(x_i) = \underset{x_j \ \in \ Convex \ Hull}{arg \ max} \frac{\left(\frac{1}{n} + x_i' x_j \right)^2}{\left(\frac{n}{n-1} - x_j' x_j \right)} \tag{1.5}$$

Which is an efficient definition in terms of computational time, so it will be used in the algorithms included in this research.

Example

To illustrate the discriminator function in the multivariate case, we present the widely known Old Faithful data set from Azzalini and Bowman (1990), considering the waiting time between eruptions and the duration of them from the geyser "Old

Faithful" in Yellowstone Park, Wyoming, USA. This data set form two groups as shown in the Figure 1.1.

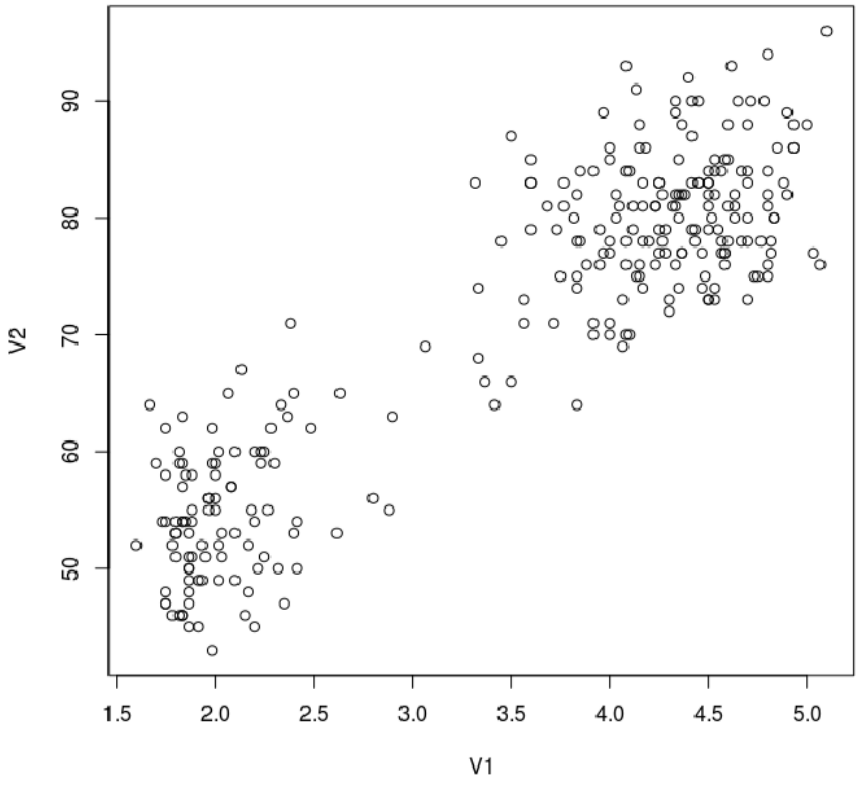

FIGURE 1.1: The Old Faithful data set

Applying the discriminator function, each data point is assigned to one discriminator following Equation (1.5) as showed in Figure 1.2, where is possible to see that the use of the discriminator function split the data into groups, assigning each point to one of the discriminators (observations 19, 58, 76, 149, 158, 161, 197, and 265) and this measure will be used in the SAR to perform the cluster analysis as we will see in the next section.

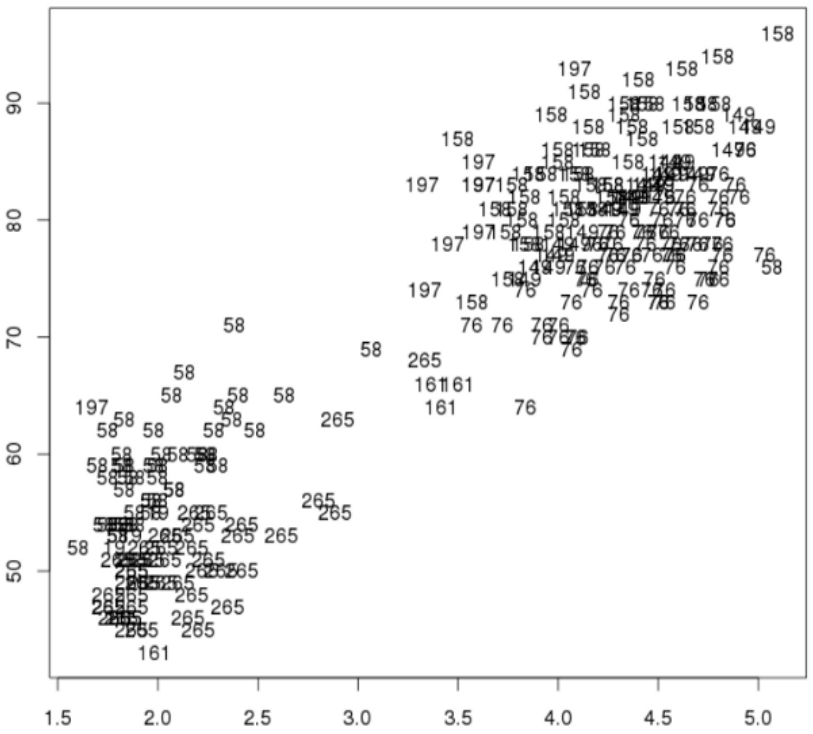

FIGURE 1.2: Discriminator function relationships, the number of the discriminator point is plotted

1.4.3 Splitting the sample

The splitting process of the SAR algorithm consists into a) identify and eliminate outliers, based on the heterogeneity measure, b) points sharing a common discriminator are put in the same group (discriminators are considered as isolated observations), c) then each group is now considered as a new sample and the procedure is continued until splitting further the sample will lead to groups that all of them are of size smaller than some minimum size n_0, and finally d) when a group cannot be split again, is called a "basic set". The minimum size is proposed as $n_0 = p + h$, where $h \geqslant 2$ and p is the number of coefficients of the fitted model.

Since the partitioning stage will tend to define many groups, it is important to have a procedure for recombining the observations after the split. The more the sample is split, the smaller the internal variability of the resulting groups, so a

process that increases the internal variability of homogeneous groups is required, incorporating new observations, but at the same time avoiding the inclusion of observations that are clearly heterogeneous with respect to the group.

1.4.4 The recombining process

The recombination process is established as follows: a) calculate $C_0(x_i)$ for each point outside the core set, b) find the nearest point x_l to the basic set, i.e. one that satisfy $C_0(x_l) = min_{x_i} C_0(x_i)$, c) if $C_0(x_l)$ is below a certain cut-off value, c_N , which depends on the size of the basic set, N, the point is incorporated into the basic set to form a new group of size $N + 1$, and the process repeats until the closest point to the group exceeds the cut-off value. Then the basic set is considered as an homogeneous group.

The cut-off values have been obtained by simulation and they are included in the SAR algorithm implementation in Matlab available under request to the authors.

After applying the recombination process to all basic sets, there are two possible situations:

a) All basic sets are increased to include the entire sample, or constitute a single partition of the sample in a set of disjoint groups and some outliers.

b) After eliminating redundancies, some enlarged basic groups overlap with others. In this case the three steps of eliminating outliers, splitting and recombination is applied again to the supplementary part of a group, treating this data as a new sample. The process continues for each basic group, creating a branch structure until the entire sample is split into several disjoint subsets. Each different form of splitting is then regarded as a Possible Data Configuration (PCD). When more than one PCD is found, the problem of choosing the best can be solved by some model selection procedure. Although other models can be used, the BIC criteria (see Section 1.3) is selected.

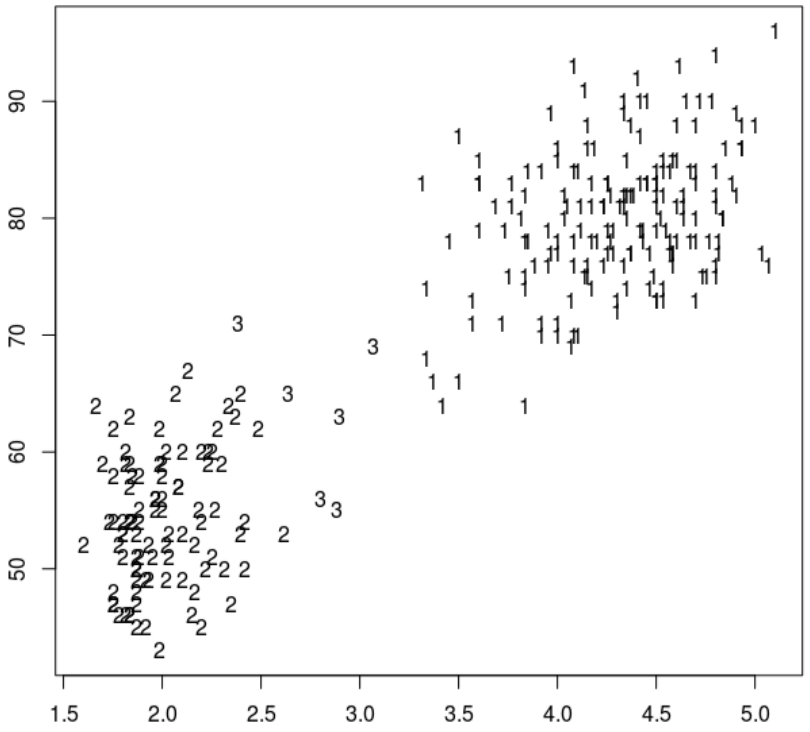

FIGURE 1.3: Result of the SAR method to the Old Faithful data set

Coming back to the Old Faithful data set (See Figure 1.1), although there are no labels in this data set, it seems to be clear the existence of two groups. After the SAR procedure is applied (see Figure 1.3), a configuration with the two main groups is founded plus a third small group of data points in the center.

1.4.5 Drawbacks and advantages of the SAR algorithm

The SAR algorithm is an efficient cluster algorithm that under certain data configurations can get good results and is a competitive alternative to classical cluster methods (Peña et al. 2004; Rodriguez 2002; Peña and Tiao 2006). The splitting process allows to identify closer observations based on the predictive probability of them given the rest of the sample, and in combination with a small minimum size for a basic group and an outlier detection method can be a efficient way to isolate observations coming from the same generator distribution.

On the other hand, the recombination process may be highly time consuming when you have a large sample size, high dimensionality and/or when too many basic groups have been detected. Moreover, the recombination process does not take into account the information obtained by the splitting process, because the observations that belong to the same basic group are regarded as isolated points during the enlargement. Third, it may fail in those cases when the groups are non linearly separable or even not separate enough: when the observations are incorporated one by one, the distance between an observation from a group i respect to a group j can be small although the structures of the groups differ.

This situation is described in figure 1.4, where we present the result of the SAR procedure applied to two bivariate normal samples of sizes $n_1 = n_2 = 100$, with means $\mu_1 = (-1, -1)$; $\mu_2 = (0, 0)$ and covariance matrices $\Sigma_1 = \begin{bmatrix} 0.25 & 0.15 \\ 0.15 & 0.25 \end{bmatrix}$; $\Sigma_2 = \begin{bmatrix} 0.25 & -0.15 \\ -0.15 & 0.25 \end{bmatrix}$. Since the limit between the two simulated groups is not clear, the SAR algorithm identifies one group and some outliers.

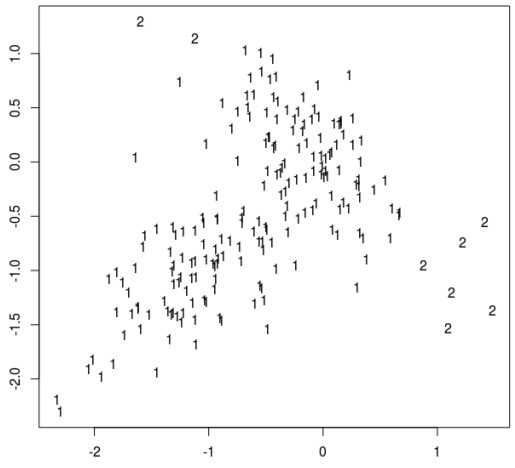

FIGURE 1.4: SAR results over two bivariate normal samples with different orientations

1.5 Conclusions

The implementation of the SAR algorithm proposed by Peña et al. (2004) deals with the model heterogeneity problem: it identifies isolated observations which not follow the same pattern as the majority in a data sample and detect possible clusters. We have seen among this chapter that the splitting part correctly identifies observations with the same structure although the recombination can be slow when the sample size is big or complex, or even fail to detect clusters when they are too close each other.

A possible improvement for these drawbacks in the recombination is to make this process not by observations, but considering each of the basic groups as a unit to recombine.

After the splitting process, each observation $x_i, i = 1, 2, ..n$ will be associated with a label vector $l = l_1, l_2, ..., l_n$, where $l_i \in \{-1, 1, 2..K\}$, being K the number of basic groups detected, and -1 the label assigned to isolated observations (i.e.) for discriminators and basic groups smaller than minimum size m_0, such that:

$$l_i = l_j \Leftrightarrow x_l(x_i) = x_l(x_j) = x_d; \forall l_i \neq -1 \tag{1.6}$$

Then for each basic group $G_g, g = 1, 2, ..K$ the C_0 measure (Equation (1.2)) is calculated for every outside observation $x_i; i = 1, 2, ..., n \notin G_g$ as described in previous section, incorporating the closest observation if the measure is under a critical value.

If we summarize the observations of the basic groups with their corresponding mean vector and covariance matrix, we obtain two clear advantages over original SAR: the process becomes more efficient in time and we keep the structure of the basic groups.

In this manner, the usual way to check if two groups come from the same population is by performing an hypothesis test like equality of means, or equality of variances test, or both at the same time.

For example, Mardia, Kent, and Bibby (1979) propose a test considering at the same time the homogeneity of means and variances through a likelihood ratio:

Let $x_{i1}, x_{i2}, ..., x_{in_i}$, be k samples i.i.d $\sim N(\mu_i, S_i); i = 1, 2, ...k$, then

$$\lambda^* = n \log |S| - \sum_i \log |S_i| - n \log |T^{-1}W| \text{ , where:}$$

$$S = n^{-1}W$$
$$S_i = \frac{1}{n_i} \sum_{j \in i} (x_j - \bar{x}_i)(x_j - \bar{x}_i)'$$
$$W = \sum_i n_i S_i; T = nS_t$$
$$S_t = \frac{1}{n} \sum_i \sum_{j \in i} (x_j - \bar{x}_i)(x_j - \bar{x}_i)'$$

this test, under $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ and $\Sigma_1 = \Sigma_2 = ... = \Sigma_k$, is distributed as:

$$\lambda^* \sim \chi^2 \left[\frac{1}{2} p(k-1)(p+3) \right] d.f.$$

.

However, after splitting in SAR the basic groups do not hold with the condition of independence, because they are not independent samples from a population, but disjoint partitions of samples. Under this situation, an alternative to recombine partitions is needed, and we will address this problem along the next chapters to improve the recombining process.

Chapter 2

Recombining by order statistics

2.1 Introduction

To split a data sample into disjoint groups by some criteria involves the definition of an order on it. In the univariate case this is clear since every partition consists in values below or above a limit number, while in the multivariate case this idea can be also extended. For this reason we decided to study the distribution of order statistics, and of linear combination of them, looking for a way to test whether two disjoint groups forms a partition (or sub partition) of the same sample. For example, we can observe the distribution of the difference between the means of the partitions.

Let $X_1, X_2, ..., X_n$ be n jointly distributed random variables, the order statistics associated with them are defined as the variables arranged in non-decreasing order such that:

$$X_{1:n} \leq X_{2:n} \leq ..., \leq X_{n:n}$$

which are also expressed in different notation as,

$$X_{(1)} \leq X_{(2)} \leq ..., \leq X_{(n)}$$

Although those variables does not need to be i.i.d. to define the corresponding order statistics, is a common assumption we will keep as a general rule.

If we split the set of the n order statistics into two groups of sizes n_1 and n_2 such that $n_1 + n_2 = n$:

$$\underbrace{X_{1:n}X_{2:n}X_{3:n}...X_{n_1:n}}_{n_1}\underbrace{X_{n_1+1:n}X_{n_1+2:n}...X_{n:n}}_{n_2}$$

then, the difference between the means of these two groups will be given by:

$$D_{21} = \bar{X}_2 – \bar{X}_1 = \frac{X_{n_1+1} + X_{n_1+2} + ... + X_n}{n_2} - \frac{X_1 + X_2 + ... + X_{n_1}}{n_1}$$

This is a linear combination of order statistics, also called "L-statistic" usually represented in this form:

$$L_n = \sum_{i=1}^{n} c_i X_{i:n} \tag{2.1}$$

where $c_1, c_2, ..., c_n$ are constants. In matrix notation, L_n is of the form:

$$L_n = c'X \tag{2.2}$$

where $c = (c_1, c_2, ..., c_n)'$ and $X = (X_{1:n}, X_{2:n}, ..., X_{n:n})$

Then is possible to write the difference D as follows:

$$D_{21} = \left(\frac{-1}{n_1}\right) X_1 + \left(\frac{-1}{n_1}\right) X_2 + ... + \left(\frac{-1}{n_1}\right) X_{n_1} + \left(\frac{1}{n_2}\right) X_{n_1+1} + ... + \left(\frac{1}{n_2}\right) X_n$$

and in this case vector of constants will be $c = \left[-\frac{1}{n_1}; -\frac{1}{n_1}... - \frac{1}{n_1}; \frac{1}{n_2}; \frac{1}{n_2}; ...\frac{1}{n_2}\right]$. If two groups come from the same population, we expect the distribution and moments of this difference will differ to the case when they come from a different distribution, and we expect to use this information to merge them.

In the first part of this chapter we will introduce some known theory about order statistics, their distributions, moments, and including original results about them in triangular distribution, a simpler and more tractable distribution which can be used to approximate a normal distribution, and in this way be applied in a wider kind of problems.

We propose a first recombination rule of univariate partitions using these previous results. Later in Subsection 2.4 of the chapter, we propose a more general approach for recombination: the use of bootstrap linear combination of order statistics. There are two ways to perform such recombination, element wise, or in a more robust alternative, using bootstrap means.

The previous techniques have limitations of being oriented to univariate problems. Thus, we extend order statistics into depth functions. There, after a brief introduction of the main measures of the depth into the data, we propose a simple recombination rule, based on the mean depth of two subpartitions.

2.2 Order statistics: theoretical introduction

The research and application of order statistics is widely spread among the literature since the classic book of David and Nagaraja (1970). A good review can be found in Arnold, Balakrishnan, and Nagaraja (2008), who claim that order statistics can be applied in problems of robust location estimates, detection of outliers, censored sampling, prediction of unlikely events, strength of materiales, reliability or quality control among others.

2.2.1 Distribution of order statistics:

Sarhan and Greenberg (1962) shows that for a sample $X_1, X_2, ..., X_n$ drawn from a distribution with cumulative distribution function F and density f , then the probability density function g of the i^{th} order statistic $X_{(i)}$ is given by:

$$g(x) = \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1}[1-F(x)]^{n-i} f(x) \tag{2.3}$$

And the corresponding cumulative distribution function $F(x_{(i)})$ is given by (David and Nagaraja 1970):

$$F(x_{(i)}) = I_{F(x)}(i, n-i+1) \tag{2.4}$$

where $I_{(x)}$ is the regularized incomplete beta function defined as:

$$I_{(x)}(a,b) = \frac{B(x,a,b)}{B(a,b)} \tag{2.5}$$

and $B(x,a,b)$ is the incomplete beta function and $B(a,b)$ the beta function. $I_{(x)}$ is also the cumulative distribution function of the beta distribution

The joint density function between two order statistics $X_{(i)}; X_{(j)} (i < j)$ is given by:

$$\begin{gathered} g(x_i, x_j) = \\ \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \left[F(x_{(i)})\right]^{i-1} \left[F(x_{(j)}) - F(x_{(i)})\right]^{j-i-1} \left[1 - F(x_{(j)})\right]^{n-j} \\ f(x_{(i)}) f(x_{(j)}) \ \forall x_i < x_j \end{gathered} \tag{2.6}$$

.

And if we choose k integers $(1 \leqslant k \leqslant n)$ such that $1 \leqslant n_1 \leqslant n_2 ... \leqslant n_k \leqslant n$, the joint probability distribution of the k order statistics $X_{(n_1)}, X_{(n_2)}, ..., X_{(n_k)}$ is given by (David and Nagaraja 1970):

$$\frac{n!}{\prod_{i=1}^{k+1}(n_i - n_{i-1} - 1)!} \prod_{i=1}^{k+1} \left[F(x_{(n_i)}) - F(x_{(n_{i-1})})\right]^{n_i - n_{i-1} - 1} \prod_{i=1}^{k+1} f(x_{(n_i)}) \tag{2.7}$$

For the domain $x_{(n_1)} \leqslant x_{(n_2)} ... \leqslant x_{(n_k)}$ and 0 otherwise, and with $n_0 = 0, n_{k+1} = n + 1; X_{(n_0)} = -\infty; X_{(n_{k+1})} = +\infty$

If we consider the k first order statistics, Equation (2.7) can be simplified as follows:

$$\frac{n!}{(n-k)!} \cdot [1 - F(x_{(k)})]^{n-k} \prod_{i=1}^{k} f(x_{(i)}) \tag{2.8}$$

and in the special case when all order statistics are considered (i.e. $k = n$), this distribution is:

$$n! \cdot \prod_{i=1}^{k} f(x_{(i)}) \tag{2.9}$$

2.2.2 Distribution of the linear combination of order statistics

Obtaining closed-form expressions for the distribution of linear combination of order statistics is generally not straightforward. For example, consider the simple case of the sum of two order statistics:

Let X,Y be a random vector with join density function $g_{X,Y}(x, y)$ Let $m(X, Y) = U, V$ be an invertible function with inverse $m^{-1}(U, V) = h(U, V) = X, Y$. If (X, Y) is continuous and h has Jacobian $J_h(u, v)$, then (U, V) has the joint density function:

$$g_{U,V}(u, v) = g_{X,Y}(h(u, v))|J_h(u, v)| \tag{2.10}$$

where $|J_h(u, v)|$ is the absolute value of the Jacobian:

$$J_h = det \begin{Bmatrix} \frac{\partial}{\partial u} h_1(u, v) & \frac{\partial}{\partial v} h_1(u, v) \\ \frac{\partial}{\partial u} h_2(u, v) & \frac{\partial}{\partial v} h_2(u, v) \end{Bmatrix} \tag{2.11}$$

Using the previous equations, let $U = U_i + U_j$ and $V = U_i$, $i < j$ by transforming density functions is possible to establish the density function of the sum of two order statistics as follows:

$$g_{U,V}(u,v) = g_{U_1,U_2}(h(u,v))|J_h(u,v)| \ ,$$

where: g_{U_1,U_2} is the joint density given by Equation (2.6), $h(u,v) = (v, u-v)$, is the inverse of the transformation function, and $|J_h(u,v)| = 1$ is the absolute value of the Jacobian of the transformation.

Developing, we get:

$$\begin{gathered} g(u,v) = \\ \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \left[F(v)\right]^{i-1} \left[F(u-v) - F(v)\right]^{j-i-1} \left[1 - F(u-v)\right]^{n-j} \\ f(v)f(u-v) \ \forall u = U_i + U_j, i < j \end{gathered} \tag{2.12}$$

The marginal distribution of the sum of two order statistics can be obtained from the previously joint distribution presented as follows:

$$g(u) = \int_v g(u,v)dv$$

Since the domain of the function requires that $x_i < x_j$, then $v < u - v \Rightarrow v < \frac{u}{2}$:

$$g(u) = \int_{-\infty}^{u/2} g(u,v)dv \tag{2.13}$$

Combining (2.13) in (2.12) is:

$$g(u)=$$

$$\int_{-\infty}^{u/2} \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \left[F(v)\right]^{i-1} \left[F(u-v)-F(v)\right]^{j-i-1} \left[1-F(u-v)\right]^{n-j}$$

$$f(v)f(u-v)dv \tag{2.14}$$

Let's consider a closer case for our problem, the distribution of the mean of two order statistics:

In this case, the transformation of variables will be $U = \frac{U_1+U_2}{2}; V = U_1$, whereby the inverse function is $h(u,v) = (v, 2u-v)$ and the Jacobian of the transformation is:

$$J_h = det \begin{Bmatrix} \frac{\partial}{\partial u}h_1(u,v) & \frac{\partial}{\partial v}h_1(u,v) \\ \frac{\partial}{\partial u}h_2(u,v) & \frac{\partial}{\partial v}h_2(u,v) \end{Bmatrix} = det \begin{Bmatrix} 0 & 1 \\ 2 & -1 \end{Bmatrix} = -2$$

therefore, $g_{U,V}(u,v) = g_{U_1,U_2}(h(u,v))|J_h(u,v)| = 2 \cdot g_{U_1,U_2}(v, 2u-v)$;

$$g(u,v)=$$

$$2 \cdot \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \left[F(v)\right]^{i-1} \left[F(2u-v)-F(v)\right]^{j-i-1} \left[1-F(2u-v)\right]^{n-j}$$

$$f(v)f(2u-v) \tag{2.15}$$

The next step is to calculate the marginal distribution, integrating with respect to v, where the domain of v might be:

$$U_1 < U_2 \Rightarrow v < 2u - v \Rightarrow 2v < 2u \Rightarrow v < u$$

$$g(u)=$$

$$2 \cdot \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \int_{-\infty}^{u} \left[F(v)\right]^{i-1} \left[F(2u-v)-F(v)\right]^{j-i-1} \left[1-F(2u-v)\right]^{n-j}$$

$$f(v)f(2u-v)dv \tag{2.16}$$

2.2.3 Order statistics moments

Let $\mu_{r:n}$ be the expected value of $X_{r:n}$, so by definition:

$$\mu_{r:n} = \int_{-\infty}^{\infty} x f_r(x) dx$$

where $f_r(x)$ is the distribution function of the r-order statistic which is given by Equation (2.3), so

$$\mu_{r:n} = \frac{n!}{(r-1)!(n-r)!} \int_{-\infty}^{\infty} x F(x)^{r-1} [1 - F(x)]^{n-r} f(x) dx \tag{2.17}$$

A bound of this value, can be obtained knowing that $0 \leqslant F(x) \leqslant 1$, so then:

$$|\mu_{r:n}| \leqslant \frac{n!}{(r-1)!(n-r)!} \int_{-\infty}^{\infty} |x| f(x) dx \tag{2.18}$$

In general the moments of product of order statistics can be obtained as follows:

$$E(X_{(n_1)}^{a_1} X_{(n_2)}^{a_2} ... X_{(n_k)}^{a_k}) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} x_{(n_1)}^{a_1} x_{(n_2)}^{a_2} ... x_{(n_k)}^{a_k} f_k(x_{(n_1)} x_{(n_2)} ... x_{(n_k)}) dx_{(n_1)} ... dx_{(n_k)}$$

$$\begin{aligned} &= \frac{n!}{\prod_{i=1}^{k+1} (n_i - n_{i-1} - 1)!} \times \\ &\int \int ... \int_{-\infty \leqslant x_{(n_1)} ... \leqslant x_{(n_k)} \leqslant +\infty} \prod_{i=1}^{k} x_{(n_i)}^{a_i} \prod_{i=1}^{k+1} \left[F(x_{(n_i)}) - F(x_{(n_{i-1})}) \right]^{n_1 - n_{i-1} - 1} \times \\ &\prod_{i=1}^{k} f(x_{(n_i)}) \prod_{i=1}^{k} dx_{(n_i)} \end{aligned} \tag{2.19}$$

So, for the particular case of variance, this will be:

$$V(X_{(i)}) = E(X_{(i)}^2) - E^2(X_{(i)})$$

$$
\begin{aligned}
&= \frac{n!}{(i-1)!(n-i)!} \times \\
&\left(\int_{-\infty}^{\infty} x^2 F^{i-1}(x) \left[1 - F(x)\right]^{n-i} dF(x) - \left\{ \int_{-\infty}^{\infty} x F^{i-1}(x) \left[1 - F(x)\right]^{n-i} dF(x) \right\}^2 \right)
\end{aligned} \tag{2.20}
$$

While the covariance between the ith and jth order statistics is:

$$
Cov(X_{(i)}, X_{(j)}) = E(X_{(i)} X_{(j)}) - E(X_{(i)}) E(X_{(j)})
$$

$$
\begin{aligned}
&= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \times \\
&\int_{-\infty}^{\infty} \int_{-\infty}^{v} u \cdot v F^{i-1}(u) [F(v) - F(u)]^{j-i-1} \cdot [1 - F(v)]^{n-j} dF(u) dF(v) \\
&- \frac{(n!)^2}{(i-1)!(j-1)!(n-i)!(n-j)!} \times \\
&\int_{-\infty}^{\infty} u F^{i-1}(u) [1 - F(u)]^{n-i} dF(u) \cdot \int_{-\infty}^{\infty} v F^{i-1}(v) [1 - F(v)]^{n-j} dF(v)
\end{aligned} \tag{2.21}
$$

2.2.4 Order statistics from normal distribution

The distribution of order statistics from continuous distributions cannot be expressed as closed-form, except for cases as exponential or uniform random variables (David and Nagaraja 1970). To calculate the moments of the standard normal distribution, several attempts has been done but only approximations are available:

A well known approximation is that given by Blom (1958), where:

$$
\mu_{r:n} = -\Phi^{-1} \left(\frac{r - \alpha}{n - 2\alpha + 1} \right)
$$

suggesting a value for $\alpha = 0.375$, which has been improved in function of r and n by Harter (1961).

Latter, Royston (1982) developed an algorithm to numerically integrate the expressions given by Equations (2.17), (2.20) and (2.21) for the standard normal case. Tables of moments for moderate sample sizes have been given by Teichroew (1956), Sarhan and Greenberg (1956) and Tietjen, Kahaner, and Beckman (1976).

The application of such expressions is limited since most of the results are not expressed in closed-form. In the next sections we will propose some alternatives for this issue.

2.2.5 Order statistics from the triangular distribution

The triangular distribution has not been deeply studied, mainly because of its simplicity and lack of applicability. Nevertheless, Kotz and van Dorp (2004) claims that triangular distribution has been revisited during the last years, and several papers based on Project Evaluation and Review Technique, PERT (Johnson 2002) and MonteCarlo simulation models (Chau 1995) have been proposed by using this distribution.

Here, we will develop some new results for order statistics from this distribution, in order to use these results to approximate those of a Normal distribution.

Let X be a random variable, so X is said to have triangular distribution if, and only if, its probability function is as follows:

$$f(x) = \left\{ \frac{4}{\theta_2^2}(\frac{1}{2}\theta_2 - |x - \theta_1|) \qquad |x| \leqslant 1 \right\}$$

where θ_1 is the mean of the distribution, and θ_2 the range.

Without loss of generality, we can assume that the distribution is centered in zero, so $\theta_1 = 0; \theta_2 = 2\theta$, and then the density function will be:

$$f(x) = \left\{ \begin{array}{cc} \frac{x+\theta}{\theta^2} & -\theta \leqslant x \leqslant 0 \\ \frac{\theta-x}{\theta^2} & 0 \leqslant x \leqslant \theta \end{array} \right\} \tag{2.22}$$

And the cumulative function is defined as:

$$F(x) = \left\{ \begin{array}{cc} 0 & x < -\theta \\ \frac{x^2+2\theta x+\theta^2}{2\theta^2} & -\theta \leqslant x \leqslant 0 \\ \frac{\theta^2+2\theta x-x^2}{2\theta^2} & 0 < x \leqslant \theta \\ 1 & x > \theta \end{array} \right\} \tag{2.23}$$

2.2.5.1 Expectation of order statistics from a triangular distribution

Proposition 2.2.1. Let $X_1, X_2, ...X_n$ be n i.i.d. triangular random variables with mean 0 and range 2θ with distribution and cumulative functions given by Equations (2.22) and (2.23). Then the expectation of the $(i : n)$ order statistic will be:

$$\begin{aligned} E(x_{i:n}) = \sqrt{2}C\theta \left[B\left(\frac{1}{2}; i+\frac{1}{2}, n-i+1\right) - B\left(\frac{1}{2}; n-i+\frac{3}{2}, i\right)\right] \\ + C\theta \left[B\left(\frac{1}{2}; n-i+1, i\right) - B\left(\frac{1}{2}; i, n-i+1\right)\right] \end{aligned} \tag{2.24}$$

where $C = \frac{n!}{(i-1)!(n-i)!} = B(i, n-i+1)^{-1} = B(n-i+1, i)^{-1}$, $B(a,b) = B(1,a,b)$ is the beta function, and $B(x,a,b)$ is the incomplete beta function defined as:

$$B(x,a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$$

Proof. Proof is given in Section A.1 □

2.2.5.2 Approximation for extreme values

Proposition 2.2.2. Let $X_1, X_2, ...X_n$ be n i.i.d. triangular random variables with mean 0 and range 2θ with distribution and cumulative functions given by Equations (2.22) and (2.23), then the expectation of the $(i:n)$ order statistic for non central values of i, can be approximated by:

$$E(x_{i:n}) \approx \begin{cases} \theta\left[\sqrt{2}\dfrac{B\left(i+\frac{1}{2}, n-i+1\right)}{B\left(i, n-i+1\right)} - 1\right] & i << \frac{n}{2} \\ \\ \theta\left[1 - \sqrt{2}\dfrac{B\left(n-i+\frac{3}{2}, i\right)}{B\left(n-i+1, i\right)}\right] & i >> \frac{n}{2} \end{cases} \tag{2.25}$$

Proof. Proof is given in apendix A.2 □

For $n = 10$, the values of $i = 1, 2, .., 5$ are shown in Table 2.1. It can be observed that in the central values the approximation is not as accurate as for the extreme values.

TABLE 2.1: True expectation and their approximation of the first five order statistics from a triangular distribution of range 2, and sample size n=10

	1	**2**	**3**	**4**	**5**
E	-0,61779	-0,42663	-0,28295	-0,16198	-0,05275
E aprox.	-0,61779	-0,42669	-0,28336	-0,16393	-0,05942
E - E aprox.	0,00000	0,00006	0,00041	0,00194	0,00667

2.2.5.3 Moments of order statistics: The triangular distribution as an approximation to the normal distribution

Let X be a random variable with symmetric triangular distribution with mean 0 and rank 2θ (minimum value $-\theta$ and maximum θ), so its distribution function is given by Equation (2.22).

It is possible to approximate a normal distribution (μ, σ^2) with a triangular distribution matching their means and variances. In the case of the given triangular distribution, $E(X) = 0$:

$$Var(X) = E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \frac{1}{\theta^2}\int_{-\theta}^{0}(x^3 + \theta x^2)\,dx + \frac{1}{\theta^2}\int_{0}^{\theta}(\theta x^2 - x^3)\,dx$$

$$E(X^2) = \frac{1}{\theta^2}\left[\frac{x^4}{4} + \frac{\theta x^3}{3}\right]_{-\theta}^{0} + \frac{1}{\theta^2}\left[\frac{\theta x^3}{3} - \frac{x^4}{4}\right]_{0}^{\theta} = \frac{1}{6}\theta^2$$

Matching the means and variances we have

$$E(X) = \mu = 0$$

and

$$Var(X) = \frac{1}{6}\theta^2 = \sigma^2$$
$$\Rightarrow \theta = \sigma\sqrt{6}$$

The equivalent normal distribution, according to the above criteria will have mean given by $\mu = 0$ and variance $\sigma^2 = \frac{1}{6}$. This approximation is presented in Figure 2.1.

The ultimate goal of this approach is to approximate the moments of order statistics from a normal distribution with those obtained from a triangular distribution. Due to the absence of an analytical equation describing the order statistics of a normal distribution we will generate $100,000$ independent random samples of size n. Once ordered from lowest to highest, we use the $100,000$ samples to obtain the mean value for each order statistic, obtaining an approximation of the actual value of its expectation, to be compared with the exact and approximated triangular order statistics.

Figure 2.2 shows the comparison for values of $n = 10$, 50, 100, and 500. As expected, because the normal distribution has heavier tails than the triangular, the latter distribution fits better the normal for non extreme values. Regarding

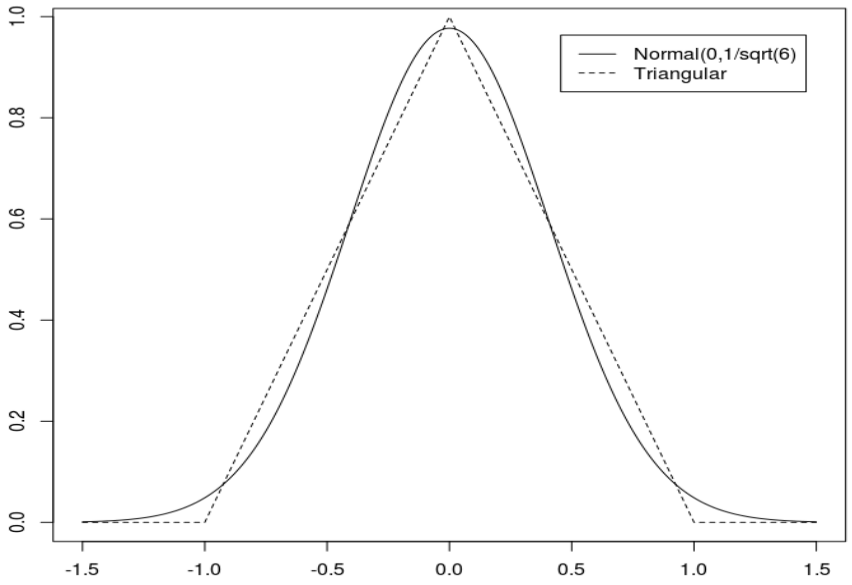

FIGURE 2.1: Approximation of a triangular distribution of range 2 to a Normal $(0; \frac{1}{6})$

the triangular approximation, it matches the full triangular moments for sizes of $n > 10$, so it can be proposed as an alternative expression for that distribution, but also as an approximation to calculate expected values of order statistics from normal distribution.

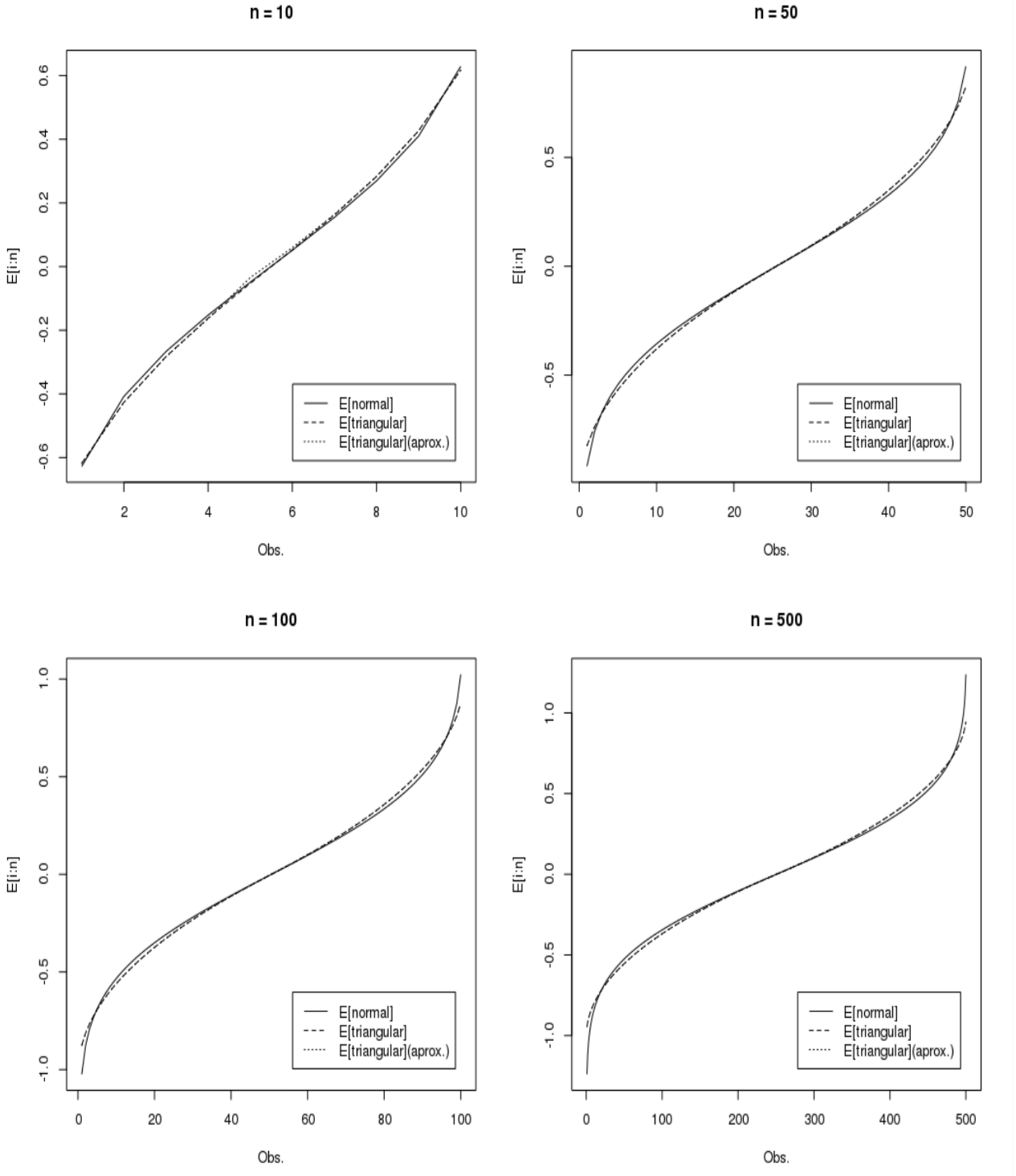

FIGURE 2.2: Expectation of order statistics from the Normal distribution for several sample sizes and then, approximating by a Triangular distribution

2.3 A first approach to recombine using confidence intervals for order statistics

A first attempt to recombine partitions using order statistics can be based on the following idea: A set of partitions should be recombined if, when they are merged, they form a sample from a given distribution. The choice of the distribution depends on the goals of the researcher, but in the simplest case, assuming that we are looking for normally distributed clusters, a set of partitions should be recombined if they form a sample from a normal distribution.

One way to evaluate it could be performing non parametric tests such as Kolmogorov-Smirnov or Shapiro-Wilk, but then we can lost some information given by the partitions, and they are not robust methods for small samples as partitions can be. An alternative is to use order statistics, by testing if the ordered observations lay into the confidence interval of the corresponding order statistic from the distribution assumed. Given that we already have some information given by the partitions, is not necessary to test all the observations, but just few point representing each partition, for example the first and last observations, i.e., the data points which "connect" the partitions. In this way, we will recombine a group of partitions if every first and last ordered observation of them lays into the corresponding confidence interval.

Formally, let $x_1, x_2, ...x_n$ be an univariate sample of size n from a certain density distribution f and cumulative distribution F, and let $x_{1:n}, x_{2:n}, ...x_{n:n}$ be the corresponding ordered sample. If we split it into k partitions of sizes $n_1, n_2, ...n_k$, such that $n_1+n_2+...+n_k = n$ following a certain criteria, we expect that when the partitions form a single sample from f, each observation will lay into the confidence interval of it corresponding order statistic.

From Equation (2.4), a confidence interval for the $X_{i:n}$ order statistic will be given by $C.I. = \{x_{lower}, x_{upper}\}$ such that:

$$I_{F(x_{lower})}(i, n-i+1) = \frac{\alpha}{2}; I_{F(x_{upper})}(i, n-i+1) = 1 - \frac{\alpha}{2} \tag{2.26}$$

Then, to recombine two given partitions, we observe if the first observations of those partitions lay in the corresponding confidence interval, assuming a distribution for the data. For the normal distribution, $F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}$.

We will illustrate this approach with two trivial examples involving normal distributions:

Example 1. One sample. We simulate a sample of size n=60 from an univariate standard normal distribution, with the sample mean $\bar{x} = 0.0656$ and standard deviation $\hat{\sigma} = 0.9104$. We sorted and split it into four subgroups of sizes $n_1 = n_2 = n_3 = n_4 = 15$,

To recombine groups 1 and 2 we observe the last element of group 1 ($x_{15:60} = -0.5558$). Considering $\alpha = 0.05$, the corresponding confidence interval will given by:

$$x_{lower} = \Phi^{-1}\left(I^{-1}(0.025, 15, 45)\right) = -1.0486$$

$$x_{upper} = \Phi^{-1}\left(I^{-1}(0.975, 15, 45)\right) = -0.3574$$

We also observe the first element of group 2 ($x_{16:60} = -0.4728$). Considering $\alpha = 0.05$, the corresponding confidence interval will given by:

$$x_{lower} = \Phi^{-1}\left(I^{-1}(0.025, 16, 45)\right) = -0.9914$$

$$x_{upper} = \Phi^{-1}\left(I^{-1}(0.975, 16, 45)\right) = -0.3092$$

, where I^{-1} and Φ^{-1} are the quantile functions of the beta and normal distributions respectively. The results of the order statistics $1, 15, 16, 30, 31, 45, 46$ and 60 are shown in Table 2.2

TABLE 2.2: Observed and confidence intervals for order statistics of a normal sample

order	**obs**	**CI(lower)**	**CI(upper)**
1	-1.9666	-3.3380	-1.5579
15	-0.5558	-1.0486	-0.3574
16	-0.4728	-0.9914	-0.3092
30	-0.0285	-0.3370	0.2951
31	0.0705	-0.2951	0.3370
45	-1.9666	-3.3380	-1.5579
46	0.7014	0.3574	1.0486
60	2.1690	1.5579	3.3380

As expected, the eight order statistics are between the values of the corresponding confidence intervals, since the recombination of all partitions forms an unique sample from a normal distribution.

Example 2. Two samples. In this case we generate two samples of 60 observations each, from an univariate normal distribution with means -2 and 2, respectively and standard deviation 1 in both cases. The combination of this two samples forms a total sample of mean $\bar{x} = -0.0249$ and standard deviation $\hat{\sigma} = 2.2388$. We sorted the total sample, and split it into eight subgroups of sizes $n_1 = n_2 = ... = n_8 = 15$

As in the previous example, we will observe if the first observation of each partition lays into the corresponding confidence interval assuming they conform a single sample of mean and standard deviation given by the full sample. The results are shown in Table 2.3

In this case, sample order statistics 30, 31, 45, 46, 75, and 76 lay outside the confidence intervals, indicating they do not conform a single sample, so they should not be recombined. A remaining problem here is that we know that all groups should not be merged but we do not know which of them should. One possible solution is to perform a backward elimination of partitions, starting with all groups and removing the last partition in each test until all order statistics lay into the confidence intervals. This toy example serves to show that is possible to combine partitions based on order statistics, although nevertheless, this simple procedure

TABLE 2.3: Observed and confidence intervals for order statistics of two normal samples

order	obs	IC(lower)	IC(upper)
1	-4.3092	-7.9187	-4.2265
15	-2.6407	-3.3010	-2.0068
16	-2.6279	-3.1982	-1.9269
30	**-2.1389**	**-2.1131**	**-1.0203**
31	**-2.0556**	**-2.0507**	**-0.9649**
45	**-1.5518**	**-1.2748**	**-0.2525**
46	**-1.4806**	**-1.2242**	**-0.2045**
60	0.0501	-0.5494	0.4529
61	0.1873	-0.5026	0.4997
75	**1.2587**	**0.1549**	**1.1745**
76	**1.3481**	**0.2028**	**1.2251**
90	1.9287	0.9153	2.0010
91	1.9659	0.9706	2.0635
105	2.5630	1.8772	3.1485
106	2.6366	1.9572	3.2514
120	5.2410	4.1769	7.8690

can be improved by considering some more robust options. We will analyse these improvements in the following sections.

2.4 Linear combination of order statistics: bootstrap approach

2.4.1 L-statistics

A L-Statistic is a linear combination of order statistics as such given by expression (2.1). This class of estimates covers a wide range of applications including the sample mean, the sample median, the range, and trimmed versions of them (Fraiman and Meloche 1999). However, as is established in the previous Subsection 2.2.2, the distribution function or moments for these expressions cannot be expressed in closed-form for main distributions, and only approximations have been

made (Stigler 1969; Balakrishnan, Charalambides, and Papadatos 2003; Rychlik 2004; Kaluszka and Okolewski 2005). Therefore, another estimation techniques have been attempted, like Bootstrap (Hutson and Ernst 2000) or Jackknife (Parr and Schucany 1982).

We will focus on the proposal of Hutson and Ernst (2000) who calculate exact bootstrap mean and variance of L-estimators based on exact bootstrap mean, variances and covariances of the whole set of order statistics from a sample.

Let

$$\hat{\mu} = (\hat{\mu}_{1:n}, \hat{\mu}_{2:n}, ..., \hat{\mu}_{n:n})'$$

and

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}^2_{1:n} & \hat{\sigma}_{12:n} & \cdots & \hat{\sigma}_{1n:n} \\ \hat{\sigma}_{21:n} & \hat{\sigma}^2_{2:n} & \cdots & \hat{\sigma}_{2n:n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{n1:n} & \hat{\sigma}_{n2:n} & \cdots & \hat{\sigma}^2_{n:n} \end{pmatrix}$$

denote the exact bootstrap mean vector of the order statistics, and the bootstrap variance-covariance matrix. Thus, the bootstrap mean and variance of the L-statistics L_n is given by Equations (2.27) and (2.28) , respectively.

$$\hat{\mu}_{L_n} = c'\hat{\mu} = \sum_{i=1}^{n} c_i \hat{\mu}_{i:n} \tag{2.27}$$

$$\hat{\sigma}^2_{L_n} = c'\hat{\Sigma}c = \sum_{i=1}^{n} c_i^2 \hat{\sigma}^2_{i:n} + 2\sum\sum_{i<j} c_i c_j \hat{\sigma}_{ij:n} \tag{2.28}$$

In order to avoid resampling, the authors use the fact that a bootstrap replication is equivalent to generate a random sample of size n from an uniform(0,1) distribution, and applying the sample quantile function $\hat{Q}(u) = F^{-1}(u) = X_{[nu]+1:n}$, with $0 < u < 1$ and $[\cdot]$ the floor function. Then, $\hat{Q}(u) = X_{r:n}$ in the region given by $\frac{i-1}{n} \leq u \leq \frac{i}{n}$, $i = 1, ..., n$

Using the uniform distribution, and the mentioned quantile function in equation (2.17), the authors obtain:

$$E^*(X_{r:n}) = \sum_{j=1}^{n} w_{j(r)} X_{j:n} \tag{2.29}$$

$$Var^*(X_{r:n}) = \sum_{j=1}^{n} w_{j(r)} (X_{j:n} - \hat{\mu}_{r:n})^2 \tag{2.30}$$

$$Cov^*(X_{r:n}, X_{s:n}) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} w_{ij(rs)} (X_{i:n} - \hat{\mu}_{r:n})(X_{j:n} - \hat{\mu}_{s:n}) + \sum_{j=1}^{n} v_{j(rs)} (X_{j:n} - \hat{\mu}_{r:n})(X_{j:n} - \hat{\mu}_{s:n}) \tag{2.31}$$

where:

$$w_{j(r)} = r \binom{n}{r} \left[B\left(\frac{j}{n}; r, n-r+1\right) - B\left(\frac{j-1}{n}; r, n-r+1\right) \right]$$

$$\begin{aligned} w_{ij(rs)} =& nCrs \sum_{k=0}^{s-r-1} \binom{s-r-1}{k} \frac{(-1)^{s-r-1-k}}{s-k-1} \left[\left(\frac{i}{n}\right)^{s-k-1} - \left(\frac{i-1}{n}\right)^{s-k-1} \right] \\ & \times \left[B\left(\frac{j}{n}; k+1, n-s+1\right) - B\left(\frac{j-1}{n}; k+1, n-s+1\right) \right] \end{aligned}$$

$$\begin{aligned} v_{j(rs)} =& nCrs \sum_{k=0}^{s-r-1} \binom{s-r-1}{k} \frac{(-1)^{s-r-1-k}}{s-k-1} \left\{ B\left(\frac{j}{n}; s, n-s+1\right) \right. \\ & - B\left(\frac{j-1}{n}; s, n-s+1\right) - \left(\frac{j-1}{n}\right)^{s-k-1} \left[B\left(\frac{j}{n}; k+1, n-s+1\right) \right. \\ & \left. \left. - B\left(\frac{j-1}{n}; k+1, n-s+1\right) \right] \right\} \end{aligned}$$

and, $B(x, a, b) = \int_0^x t^{a-1}(1-t)^{b-1} dt$ is the incomplete beta function, and $nCrs = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$

Proceeding by this way, the error due to bootstrapping resampling is eliminated, and the expectation and variance of any linear combination of order statistics can be obtained. By using this approach, we can construct bootstrap comparisons to help detecting whether two partitions come from the same distribution and should be (or not) merged. We will address this problem in the next subsection.

2.4.2 The bootstrap element-wise comparison

Let $x_1, x_2, ..., x_n$ be a sample from a certain distribution later ordered and splitted in two groups of sizes n_1 and n_2, with $n_1 + n_2 = n$, such that $x_{1:n}, x_{2:n}, ..., x_{n_1:n} \in$ group 1 and $x_{n_1+1:n}, x_{n_1+2:n}, ..., x_{n:n} \in$ group 2. In this section we propose the use of bootstrap expectations of order statistics to check if this two groups should be recombined.

Consider the first observation of the second group: $x_{n_1+1:n}$, this data point can also be expressed as the first order statistic of the second sample, i.e. $x_{1:n2}$, then, the main idea is to compare the bootstrap expectation of this element under these two definitions. When the groups are close each other, we expect that $E(X_{n_1+1:n}) \sim E(X_{1:n_2})$, since if the two groups are well separated, the inclusion of the first group in the total boostrap will move the expectation away from the first element of the second group.

As an example, consider a sample of size $n = 20$ generated from a standard univariate normal distribution. After ordering it and splitting it arbitrarily into two groups, leaving the negative numbers in one group and the positive in aonther, we obtain the groups given by Table 2.4

TABLE 2.4: Ordered sample from an univariate standard normal distribution

	1	2	3	4	5	6	7	8	9	10
G1	-1.29	-1.26	-0.91	-0.90	-0.64	-0.38	-0.30	-0.16	-0.03	-0.01
G2	0.10	0.19	0.28	0.44	0.86	1.03	1.24	1.24	1.43	2.35

Table 2.5 shows 10 of 500 bootstrap samples taken from the second group, and the histogram of the first element of these samples is given by Figure 2.3. Equivalently, Table 2.6 shows the first 11 elements of 5 from 500 bootstrap samples taken now from the the entire sample, and the corresponding distribution of the 11th element is shown in Figure 2.4. Is expected that these two graphs differ since the bootstrap of the second group is limited by its first element, nevertheless, we are interested in evaluate the difference between the two expectations, which empirical distribution is given by Figure 2.5.

In base of this simple example we can see that the difference is close to zero (0.0438), but this procedure seems to be not robust enough, because is based in the bootstrap of only one element. Following the same structure is also possible to consider the difference between the means of the second group (bootstrapping only from it) and the second part obtained bootstrapping from the entire sample. We present this approach in the next section.

TABLE 2.5: Ten bootstrap samples obtained from the second group of the data from Table 2.4

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	...
1	0.10	0.10	0.19	0.19	0.10	0.10	0.10	0.19	0.28	0.10	
2	0.86	0.19	0.19	0.19	0.10	0.28	0.19	0.44	0.44	0.10	
3	0.86	0.28	0.19	0.86	0.19	0.28	0.28	0.44	0.86	0.19	
4	0.86	0.28	0.28	0.86	0.19	0.44	0.44	0.44	0.86	0.28	
5	0.86	0.86	0.28	0.86	0.19	0.86	0.86	0.44	0.86	0.28	
6	1.03	0.86	1.24	1.43	0.44	1.43	1.24	0.86	1.03	1.24	
7	1.03	1.24	1.24	1.43	0.86	1.43	1.24	1.03	1.24	1.24	
8	1.24	1.24	1.24	1.43	1.03	2.35	1.24	1.03	1.24	1.24	
9	1.24	1.24	1.24	2.35	1.24	2.35	1.24	1.03	1.43	2.35	
10	2.35	2.35	2.35	2.35	1.43	2.35	1.24	1.24	2.35	2.35	

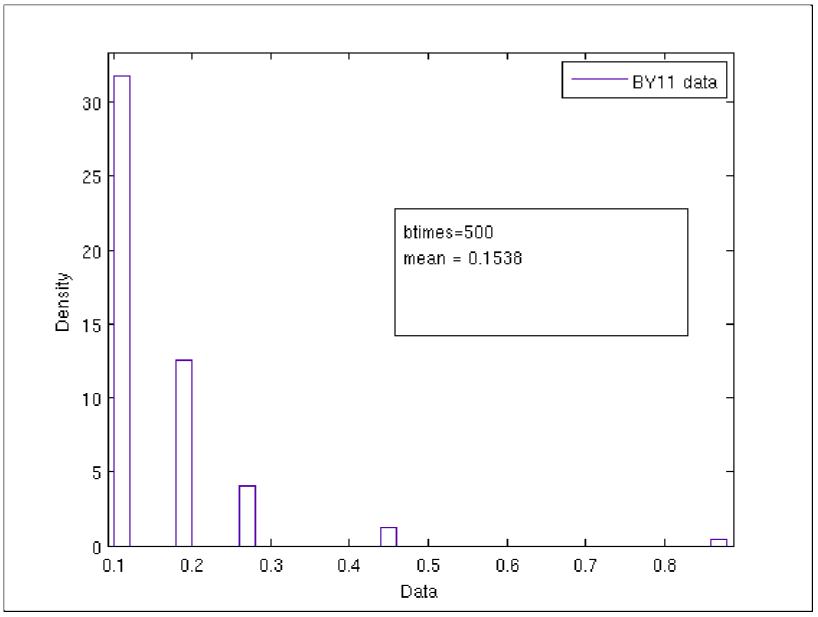

FIGURE 2.3: Distribution of the first element of the bootstrap samples given by Table 2.5

TABLE 2.6: Bootstrap samples obtained from the entire sample from a N(0,1)

	B1	B2	B3	B4	B5	...
1	-1,2926	-1,2926	-1,2926	-1,2649	-0,9082	...
2	-1,2649	-0,9082	-0,8987	-0,9082	-0,8987	...
3	-1,2649	-0,8987	-0,3797	-0,9082	-0,8987	...
4	-0,3797	-0,8987	-0,2996	-0,8987	-0,6361	...
5	-0,2996	-0,6361	-0,2996	-0,3797	-0,6361	...
6	-0,1624	-0,6361	-0,1624	-0,3797	-0,3797	...
7	-0,1624	-0,3797	-0,1624	-0,3797	-0,2996	...
8	-0,1624	-0,2996	-0,0346	-0,2996	-0,2996	...
9	-0,1624	-0,1624	0,1028	-0,2996	-0,2996	...
10	-0,0119	-0,1624	0,1028	-0,0346	-0,0346	...
11	-0,0119	-0,1624	0,2769	0,1924	-0,0119	...
...	...	...	...	...	...	...

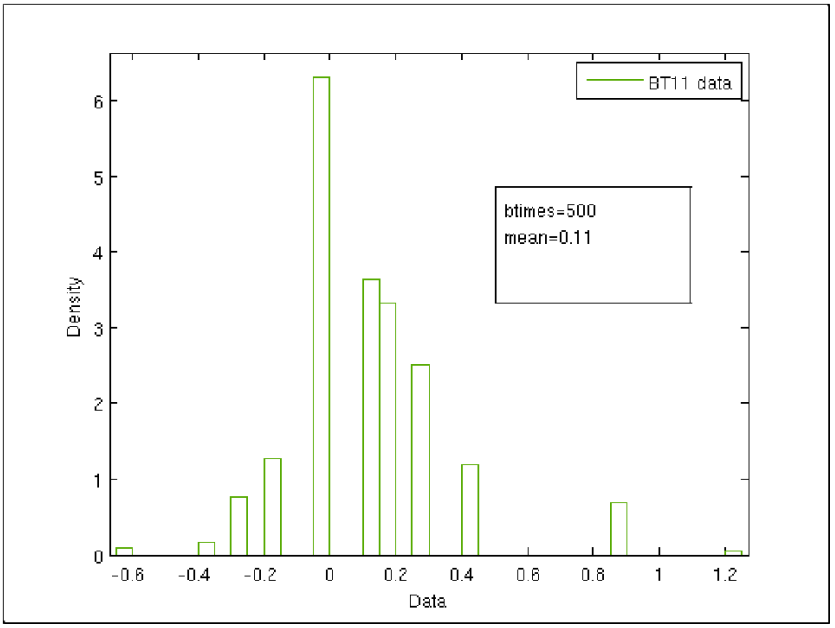

FIGURE 2.4: Distribution of the 11^{th} element of the bootstrap samples given by Table 2.6

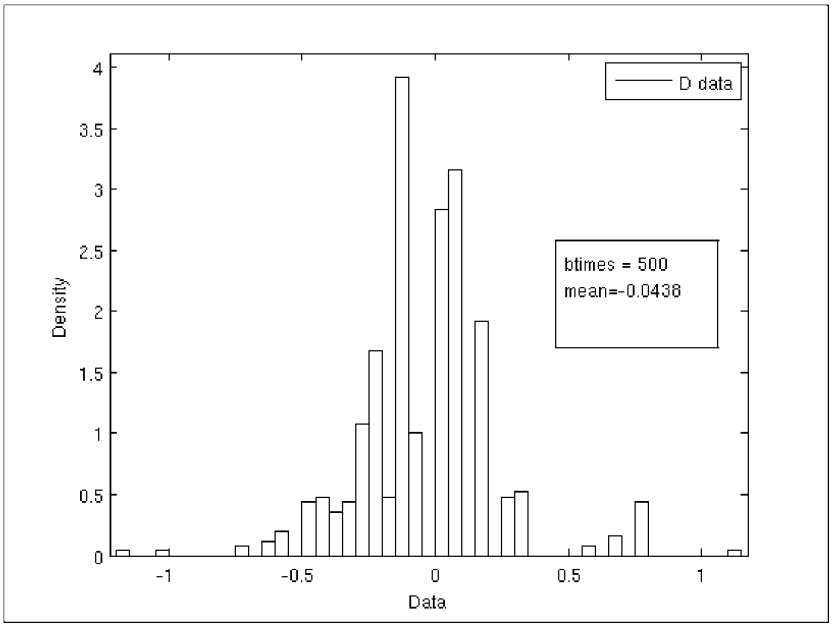

FIGURE 2.5: Distribution of the difference between the bootstrap first element of group 2 and 11^{th} of the total sample.

2.4.3 The bootstrap mean comparison

Considering again a sample $x_{1:n}, x_{2:n}, ..., x_{n_1:n}, x_{n_1+1:n}, x_{n_1+2:n}, ..., x_{n:n}$. We bootstrap b times from the second group $\left(x_{(n1+1)}, x_{(n1+2)}, ..., x_{(n2)}\right)$ and b times from

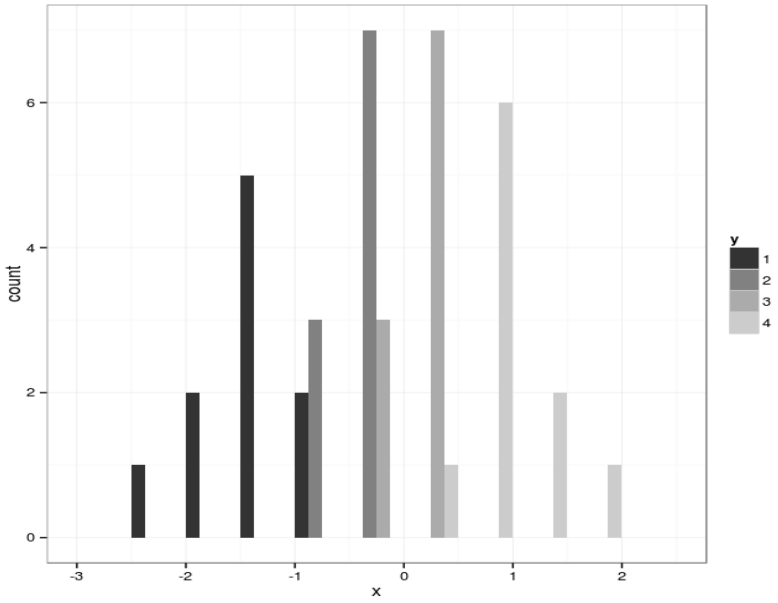

FIGURE 2.6: Partition methodology of a Normal Distribution in fourgroups

the entire sample. In order to have a more robust bootstrap comparison, we propose for each bootstrap resample to calculate the difference between the mean of the bootstrap of the second group and the mean of the last n_2 observations obtained from bootstrapping the entire sample:

$$\bar{X}_{(1:n_2)} - \bar{X}_{(n1+1:n1+n2)}$$

As an example, we generate 1000 (sorted) samples from a standard Normal distribution ($n = 40$), and then split it into 4 groups of 10 observations each from lowest to highest values in the way shown by Figure 2.6.

Taking into account the first two partitions, groups 1 and 2, from the left side of the distribution. These two partitions are adjacent each other so they should be recombined for any proposed procedure, even if together they do not form an independent random sample from a normal distribution.

The bootstrap distribution of the difference between the groups 1 and 2 is shown in Figure 2.7

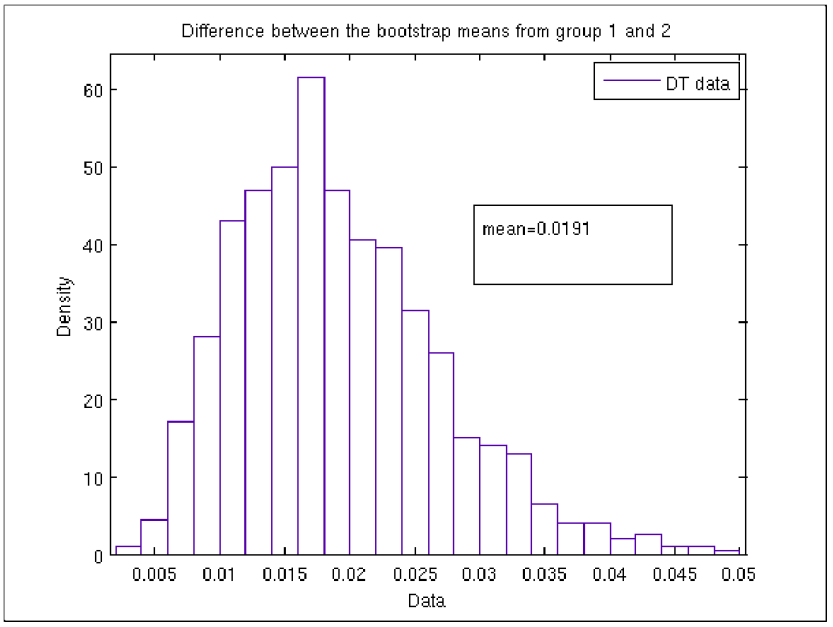

FIGURE 2.7: Distribution of the difference between bootstrap means 1 and 2.

Ideally we will recombine the partitions when there is no difference between the bootstrap order statistics, but because of the construction of the bootstrap methodology applied here, it is impossible for this difference to be centred at zero (one group is always greater than other), but if both groups are close enough, forming part of the same split, the expectation of the difference of the bootstrap means will be small (In this case, the mean is 0.0191).

Consider now the two tales of the distribution (groups 1 and 4), although they are part of the same normally distributed sample, they should not be directly recombined without taking into account the partitions in between. Figure 2.8 shows the bootstrap distribution of the difference between the two means, which in this case is bigger than the obtained in the previous example (groups 1 and 2). This information can be used to build merging rules, leading to a recombination when the difference between the mean of the order statistics from the two groups is small enough.

In order to extend the previous examples, now we simulate 1.000 samples from a standard normal distribution of size $n = 100$, and then split them into 2, 3, 4, and 5 groups each time. We calculate the bootstrap expectations for the mean of group 2 and the mean of second part of the entire sample, with the methodology of Hutson

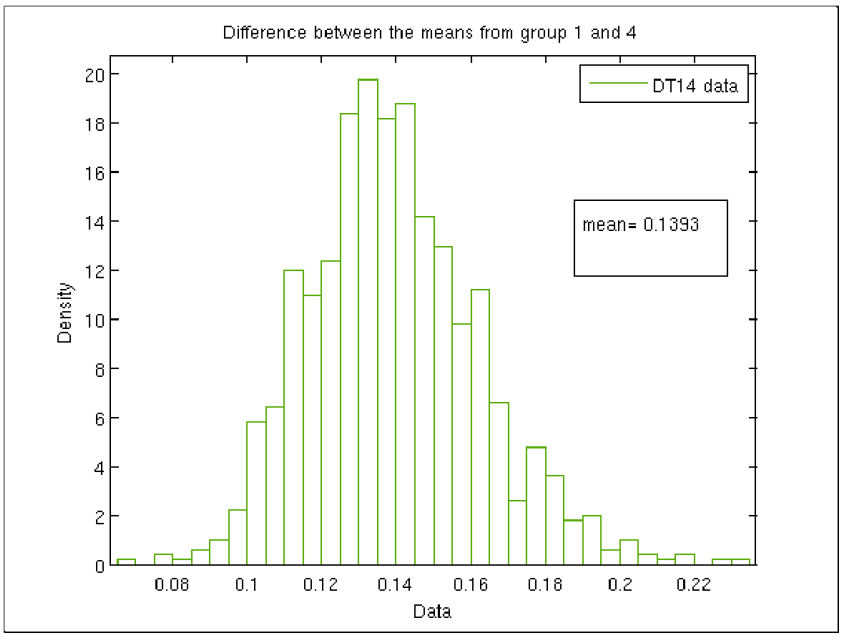

FIGURE 2.8: Distribution of the difference between bootstrap means 1 and 4.

and Ernst (2000) presented above, in order to avoid generating bootstrap samples and their resampling error. Table 2.7 show the means and standard deviation of the 1000 samples for each splitting process.

Then, when the groups are close each other and constitute a partition, we get that the expectation of the linear combination of order statistics is smaller than 0.01 in all cases, and this does not depend on what part of the distribution the recombination is evaluated, either over the tales (like in groups 1-2, 4-5, for the 5 groups example) or in the center (like in 2-3 for the 4 groups example). From this results it is possible to construct cut-offs which allow to recombine previously split data set, where there is no information about what part of the distribution the split group come from.

Considering the objectives of this research, we will not delve into this approach given its limited application because: a) It is necessary to simulate all possible combinations of sizes and groups, and b) It is a method that can not be directly applied to multivariate data. However, it is important to remark that it can be a promising starting point for future research in the field of order statistics, since certainly allow to identify areas within the data which can belong to the same distribution.

TABLE 2.7: Means of the difference between bootstrap expectations of split samples. Standard deviation in parenthesis

groups	2	3	4	5
1 2	**0.0062**	**0.0069**	**0.0079**	**0.0091**
	(0.0018)	**(0.0023)**	**(0.0028)**	**(0.0032)**
1 3		0.0488	0.0453	0.0445
		(0.0064)	(0.0068)	(0.0078)
2 3		**0.0070**	**0.0063**	**0.0064**
		(0.0023)	**(0.0022)**	**(0.0024)**
1 4			0.0838	0.0762
			(0.0088)	(0.0091)
2 4			0.0448	0.0381
			(0.0068)	(0.0066)
3 4			**0.0079**	**0.0063**
			(0.0026)	**(0.0023)**
1 5				0.1140
				(0.0114)
2 5				0.0759
				(0.0094)
3 5				0.0441
				(0.0079)
4 5				**0.0090**
				(0.0033)

2.5 Recombining by depth functions

2.5.1 Depth functions as order statistics extensions

Given the limitations of the previous measures intended for a univariate framework, is natural to look for an extension of these proposals to multivariate data. The first drawback is that $\mathbb{R}^p$ is not an ordered set, so order statistics cannot be defined in the same way as in the univariate case.

Several attempts have been done in order to define multivariate order statistics, among those are: to use an univariate measure as in Bairamov and Gebizlioglu (1997), calculate marginal order statistics as in Arnold, Castillo, and Sarabia (2009b), or consider them concomitants of some continuous univariate random variable as in Arnold, Castillo, and Sarabia (2009a).

Nevertheless, there is agreement in the literature that the concept of depth is a more appropriate equivalent to order in a multivariate dimension. The advantage of the use of depth is given by a well established theory, though is still under development (Fraiman and Meloche 1999). In general, depth is a measure of "centrality" of a multivariate data point respect to a given data set (Ding, Dang, Peng, and Wilkins 2007). In this way is possible to define an order or ranking of the data, being the median the "deepest" point and decreasing the depth as the data points are more distant from the center.

Besides defining a median in $\mathbb{R}^p$ and providing a way to measure centrality of the rest of the points, is possible to use depth functions to extend useful functions from univariate to multivariate as outlier detection, quantile functions, sign and rank functions, skewness, and kurtosis (Serfling 2012). Some recent applications of the depth include the analysis of functional data (López-Pintado and Romo 2009), and microarray data (López-Pintado, Romo, and Torrente 2010).

Formally, and following Cascos, López, and Romo (2011) and Dyckerhoff (2004), given a probability distribution P in $\mathbb{R}^p$, a depth function is a bounded function $D(;P) : \mathbb{R}^p \to [0,1]$ assigning to each point of $\mathbb{R}^p$ a degree of centrality over P. A depth function should hold the following properties:

- *Affine invariance.* $D(Ax + b; P_{AX+b}) = D(x; P_X)$ for every non-singular $A \in R^{dxd}$ and $b \in R^d$;
- *Vanishes at infinity.* $D(x; P) \longrightarrow 0$ if $\| x \| \to \infty$;
- *Upper semicontinuity.* $x \in R^d : D(x; P) \geqslant \alpha$ is closed;
- *Monotonicity relative to deepest point.* $D(x; P) \leqslant D(\theta + \lambda(x - \theta); P)$for $\theta = \text{argmax}_x D(x; P)$ and $\theta \leqslant \lambda \leqslant 1$;
- *Quasiconcavity.* $D(\lambda x + (1-\lambda)y; P) \geqslant \min\{D(x; P), D(y; P)\}$ for $\theta \leqslant \lambda \leqslant 1$

Let x be a vector in $\mathbb{R}^p$, and $X_1, X_2, ...X_n$ a data set with a corresponding F distribution function. Several definitions of depth of x respect the data set have been proposed, being the following the most widely used:

Halfspace depth (Tukey 1975; Donoho and Gasko 1992). Also known as Tukey's depth at x is defined as

$$\mathrm{TD}(x) = \inf_H \{F_n(H) : x \in H\},$$

where H is a half space; and F_n the empirical distribution.

Mahalanobis Depth (Mahalanobis 1936). Is defined as the inverse of the Mahalanobis distance as:

$$\mathrm{MD}_n(x) = \frac{1}{1 + (x - \bar{X}_n)^T \Sigma^{-1} (x - \bar{X}_n)}$$

where $\bar{X}_n$ and Σ are the sample mean and covariance matrix respectively.

Simplicial Depth (Liu and Singh 1993; Liu 1988, 1990). The simplicial depth at x counts in how many closed simplex with vertices in the sample is the point x.

$$\mathrm{SD}(x) = P_{Fn(x)} \in S[X_1, ..., X_{p+1}]), \tag{2.32}$$

where $S[X_1, ..., X_{p+1}]$ is the closed simplex with vertices $X_1, ..., X_{p+1}$, and F_n the empirical distribution.

Spatial depth (Brown 1983; Chaudhuri 1996). The sample spatial depth is defined as:

$$\mathrm{SPD}(x) = 1 - \left\| \frac{1}{n} \sum_{i=1}^{n} S(x - X_i) \right\|$$

where $S(x) = x/\|x\|$is the spatial sign function ($S(0) = 0$) with Euclidean norm $\|\|$.

Projection depth (Zuo and Serfling 2000). Is based on the biggest discrepancy between a one-dimensional projection x and the median point of the same projection applied to the data set X.

$$\mathrm{PD}(x) = \left(1 + \sup_{u \in \mathbb{R}^p} \frac{< x, u > - Me(< X, u >)}{MAD(< X, u >)} \right)^{-1}$$

where Me is the median, and MAD is the median absolute deviation.

Oja depth (Oja 1983). Is constructed from the expected volume of a simplex with a fixed vertex x and the rest random. The sample version is calculated as the average volume of the simplices constructed from all subsets of p different observations from the sample

$$\mathrm{OD}(x) = \left(1 + \frac{\mathbf{E}\mathrm{Vol}_d(\mathrm{co}\{x, X_1, ..., X_p\})}{\sqrt{\det\Sigma}}\right)^{-1}$$

where *co* is a closed simplex.

Other depths are majority depth (Singh 1991), zonoid depth (Koshevoy and Mosler 1997), generalized Tukey depth (Zhang 2002), and L1 depth (Vardi and Zhang 2000). Good summaries of some of these depths are given by Zuo and Serfling (2000) and Cascos et al. (2011).

The simplicial depth (Liu and Singh 1993; Liu 1988, 1990), defined in Equation (2.32), calculates the probability of a given point x of being inside a simplex whose vertices are randomly chosen from the rest of the sample. In this way, the more centred data point is, the highest probability of be surrounded by other points and have a high depth.

As an example, consider a sample of size $n = 200$ taken from a bivariate normal distribution of mean zero and covariance matrix identity, as those given by Figure 2.9. In general, from a data set of size n in dimension p we can generate $\binom{n}{p+1}$ simplices, so for this data since n=200 and p=2 we get 1,313,400 triangles. A random sample of 100 of those triangles is plotted over the data set to visualize the depth measure in Figure 2.10, where we observe that points in the center are covered by more triangles than those points over the borders.

Now consider two separated samples of sizes $n_1 = n_2 = 100$ generated from a bivariate normal distribution with means $\mu_1 = (-3, -3)$ and $\mu_2 = (3, 3)$ and both with identity covariance matrix. Figure 2.11 shows the corresponding total

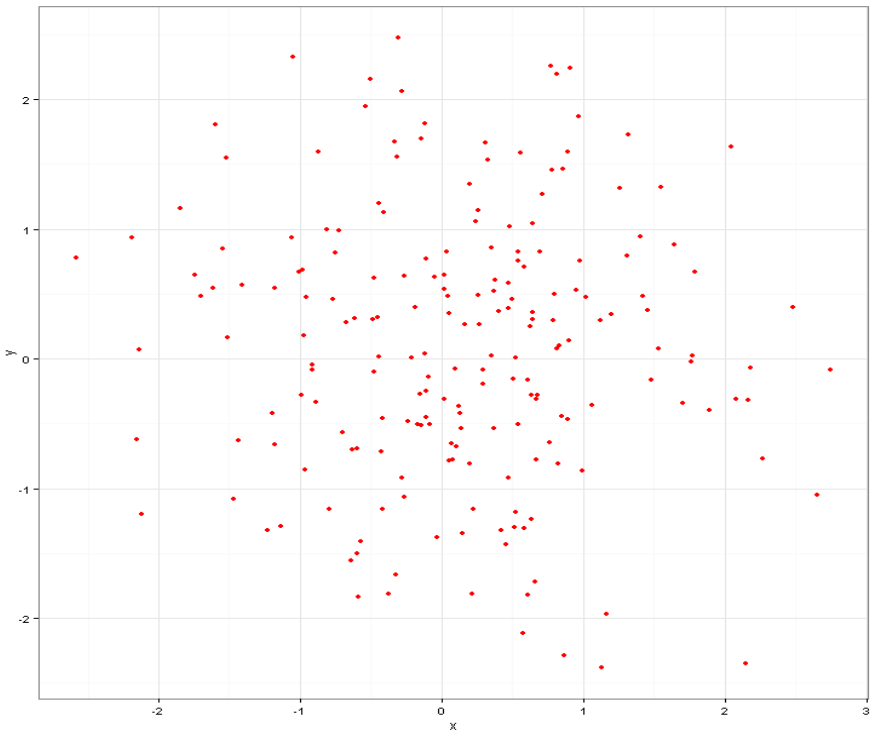

FIGURE 2.9: A bivariate standard normal sample

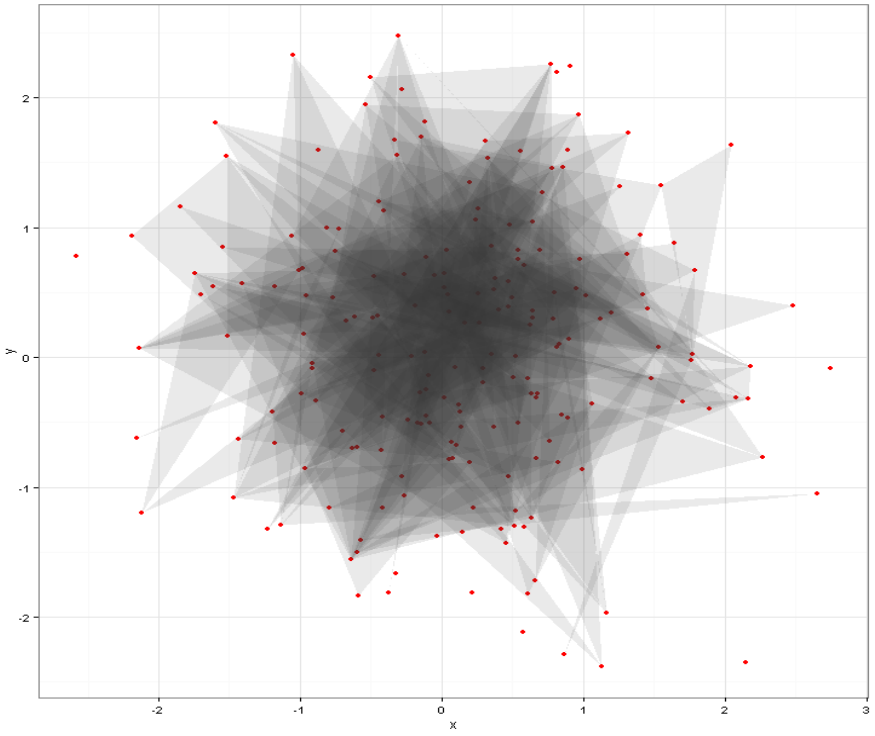

FIGURE 2.10: Simplicial depth over a bivariate standard normal sample

triangles. Now the depth structure differ since the deepest zone is concentrated in an area with no data points.

It is clear that the depth is a good measure of centrality, similar to order statistics in one dimension, so it can be also used in a context of a recombining strategy, since similar portions of the dataset have similar values of the depth.

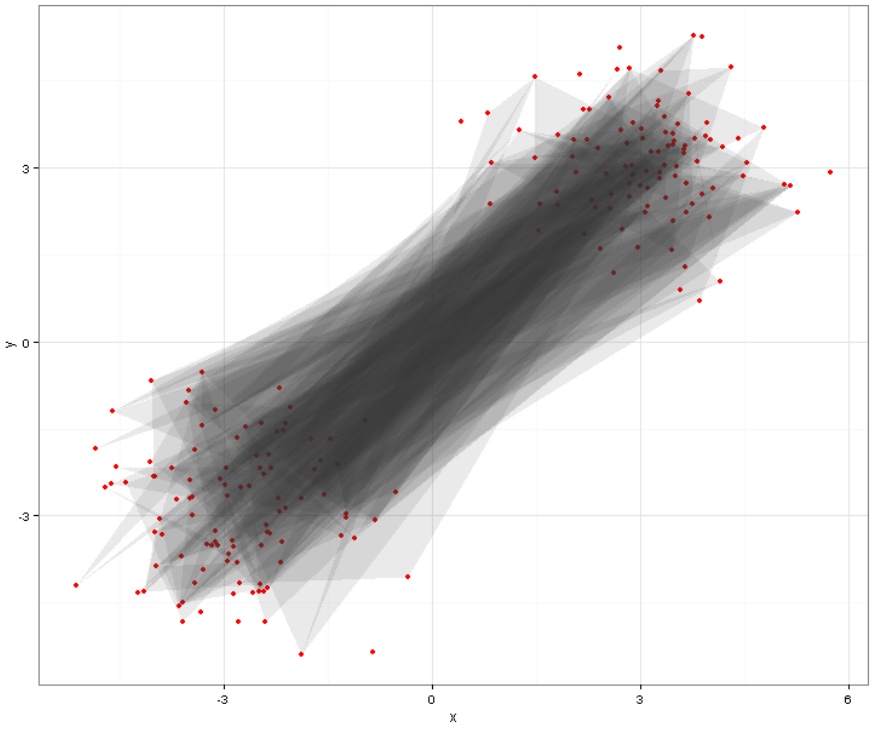

FIGURE 2.11: "Total" simplicial depth over two bivariate standard normal samples

2.5.2 A simple recombination rule based on simplicial depth.

When two partitions are close to each other, they should also have similar depth values in comparison to the total data sample. Then, a simple test to recombine those partitions can be a t-test or a Wilcoxon ranked test for homogeneity of means. Notice that both tests assume observations are independent, which as we remarked in Chapter 1, cannot be hold in those partitions. Nevertheless, they can be useful in order to detect how similar values of depth the partitions have.

Formally, let $(X, l) = (x_1, l_1), (x_2, l_2), ..., (x_n, l_n)$ a sample of n observations from an unknown distribution f (possibly a mixture), and the corresponding labels such that $l_i \in \{1, 2, ..., k\}$ assigning each observation x_i to one of $k \leq n$ possible partitions, predefined by a certain partition procedure.

Let $X_i = (x_1, l_i), (x_2, l_i), ..., (x_{n1}, l_i)$ and $X_j = (x_1, l_j), (x_2, l_j), ..., (x_{n1}, l_j)$ two of the partitions. We will assign these partitions to different clusters by testing

$$H_0 : \mu_{di} = \mu_{dj}$$

$$H_1 : \mu_{di} \neq \mu_{dj}$$

where μ_{di} is the mean of the depth of the partition i, estimated by the sample mean $\bar{d}_i = \frac{1}{ni} \sum_{x \in X_i} d(x, X)$ and $d(x, X)$ is the depth of the observation x respect to the full sample X. In particular, we propose the use of the Liu depth (simplicial) and the Wilcoxon test for the means. When the corresponding p-value $\leq \alpha$ we reject the null and the two partitions should be assigned to different clusters.

2.5.3 Example

Recall the geyser data set introduced in Section 1.4.2, Figure 1.1. After applying it the discriminator function and lately assigning the isolated observations to the closer groups using the Mahalanobis distance, we get the total sample split into five subpartitions as shown in Figure 2.12.

After splitting, we calculate the simplicial depth for each observation relative to the full data set. The depths were calculated using the R package 'depth' (Genest, Masse, and Plante 2012), although also for R are available the packages 'localdepth' (Agostinelli and Romanazzi 2009), and 'depthTools' (Lopez-Pintado and Torrente 2013). Starting from the biggest group (labeled 1), we apply hierarchically the Wilcoxon test to compare the group 1 with the rest. The corresponding p-values are shown in table 2.8, where the group 1 is merged with the partitions 2 and 3. Then we get a p-value smaller than $\alpha = 0,01$ separating the upper right

cluster of the geyser sample. Finally the groups 4 and 5 are not split, so they conform the same cluster, leading to the final data configuration shown in Figure 2.13.

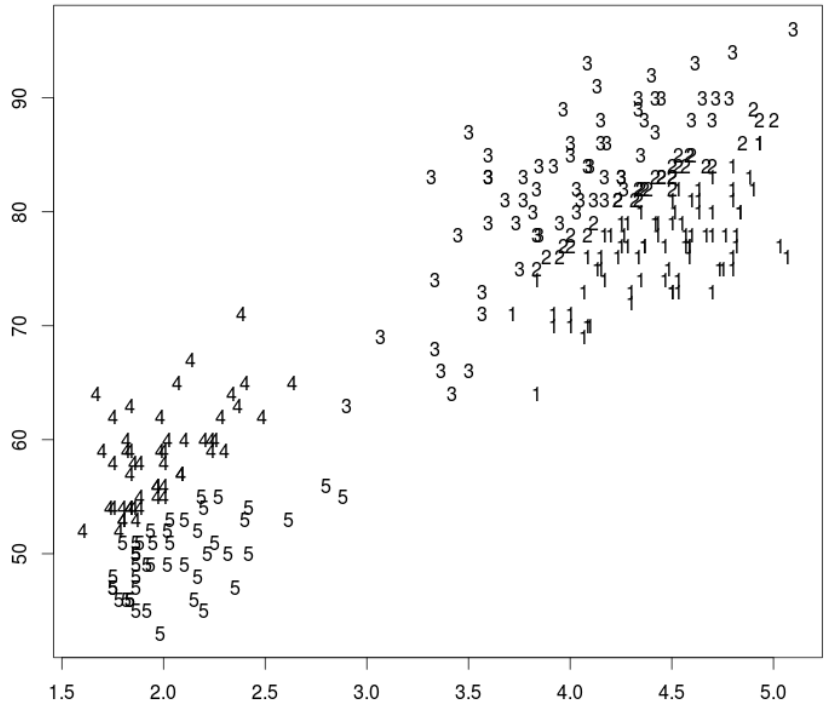

FIGURE 2.12: Five partitions from the geyser data

TABLE 2.8: Hierarchical recombination test based on depth, over five partitions of the geyser data

groups	p-value
1 - 2	0,0140
12-3	0,0911
123-4	0,0003
4 - 5	0,2340

2.6 Conclusions

When trying to develop recombining procedures, the election of order statistics seems to be natural: since the partitions of the data are mutually exclusive with

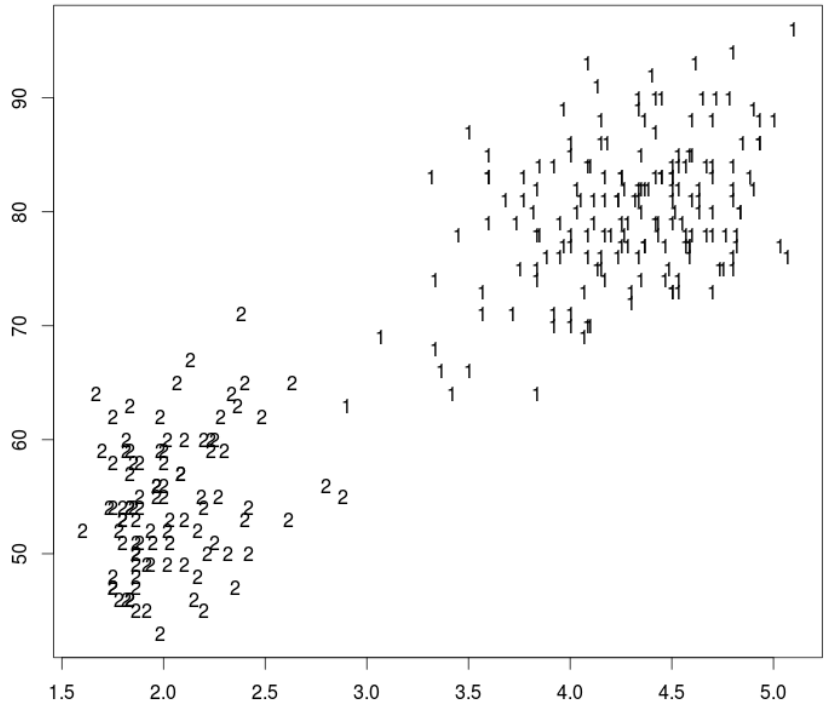

FIGURE 2.13: Cluster results of the geyser data based on depth recombination

no overlaps, they also define an order on the data. Nevertheless, along the first part of the chapter we showed that the use of order statistics often leads to non tractable expressions, so another alternatives should be considered.

One of them is the use of bootstrap resampling methods, considering that if two partitions are adjacent, the corresponding last and first order statistics should be similar. In addition, we propose a more robust alternative by comparing the bootstrap mean of the two partitions.

As we mentioned before, some of the advantages of using a bootstrap methodology is that large sizes of data samples are not needed, and by using the exact moments proposed by Hutson and Ernst (2000) the resampling error can be avoided.

In terms of drawbacks, the application of these methodologies over multivariate samples is limited, since although bootstrap methods are easily implemented in p>2 dimensions, in order to obtain the order statistics moments is necessary to

implement reduction of dimensionality through the use of projections as in Peña and Prieto (2001), or defining such statistics in a multivariate framework.

A first attempt to solve this problem is given in Section 2.4 where depth functions are proposed as an extension of the order statistics. The underlying idea is that two partitions should be recombined if they have similar depth values, respect to the total sample. Results shown that this approach can be a useful way to form final clusters, nevertheless, is not a robust methodology since the procedure strongly depends on the partition method: for example if we split the data of the Figure 2.11 in the two original samples, both will have similar depth mean leading to a wrong recombination. Along the next chapters we will explore other alternatives to overcome these limitations.

Chapter 3

Recombination by means of unimodality tests

3.1 Introduction

In this Chapter we introduce a non parametric approach to merge partitions by checking whether the data can be assumed as unimodal or not. As we have seen in the introduction of Chapter 1, this is a natural way to identify the presence of clusters, understanding each of them as a mode surrounded by a density and separated enough from other modes, if they exist. Unimodality is well defined for univariate data sets but these techniques can be extended to multivariate analysis by: a) projecting the data into one dimension and then evaluating the unimodality, or b) choosing one of the possible unimodality definitions and techniques for multivariate data.

A distribution F is defined as "unimodal", if F is convex for $x < m$ and concave for $x > m$, where m is the mode. Under this definition is clear that the normal, the student or the chi-squared distributions are unimodal, but also the uniform(a,b) distribution is considered unimodal under this definition, given that m can be any value in $[a, b]$.

Hartigan and Hartigan (1985), introduce the "dip test" to detect the presence of one or multiple modes into the data. Given the empirical distribution, the dip statistic computes the maximum difference between that distribution and an unimodal distribution function in the following way:

Let $x_1, x_2, ..., x_n$be a set of univariate data coming from a density function $f(x)$, and $F_n(x)$ be the sample empirical distribution function. Let $H(x)$ be the closest unimodal c.d.f. respect to the empirical distribution, then the DIP statistic is given by:

$$DIP = \sup_x |F_n(x) - H(x)| \tag{3.1}$$

Although Bickel and Fan (1996) show that the non-parametric maximum likelihood estimate of the closest unimodal cdf, given the mode location m_0, is the greatest convex minorant of F_n on $(-\infty, m_0]$ and the least concave majorant on $[m_0, -\infty)$ (Tantrum, Murua, and Stuetzle 2003), the authors of the test propose the use of an uniform distribution to obtain a critical value to compare the statistic. They claim that the dip is asymptotically larger for the uniform than for any unimodal distribution with exponentially decreasing tails, so this choice implies being very conservative in the assumption of the underlying distribution of the data.

Cluster methods like M-clust (Fraley and Raftery 1998), model the underlying distribution of the data by a mixture of normal distributions. The parameters are estimated by the EM algorithm, while the Bayesian Information Criteria (BIC) is used to decide the number of groups, by estimating the number of components of the mixture which maximize the likelihood, penalized by the number of estimated parameters. (See Chapter 1 and references therein for more details)

The problem with this kind of estimation arises when the true data is not a mixture of normals, and other distribution can fit better, or when the concept of "cluster" is not equivalent with the number of components of the mixture. For example,

when a cluster is defined by finding gaps in the density, a mixture of normals can be not appropriate to define the number of groups.

The dip test has been used by Tantrum et al. (2003) as a tool to identify whether a mixture of normal distributions overestimates the real number of clusters in a sample. They propose an algorithm for pruning the cluster tree generated by the mixture model chosen by the Model Based Clustering. It then progressively merges clusters that seems to be unimodal by using the dip test. A similar approach to Tantrum et al. (2003) is proposed by Ahmed and Walther (2012) who project multivariate data on its principal curves and then apply Silverman's multimodality test (Silverman 1981) to the resulting univariate sample. Other methods specifically designed to merge Gaussian components will be reviewed in Chapter 4, and a recent comprehensive literature review about this topic can be found in Hennig (2010a).

3.2 Recombining with the dip test

Given a data sample $x_1, x_2, ...x_n$ of n i.i.d. observations coming from an unknown distribution function, we apply the discriminator function to classify the observations into $k \leq n$ partitions. We split the sample in the same way as the SAR algorithm (Peña et al. 2004), where the discriminator is the observation which appears as most discrepant with respect to the rest of the data set when the discriminator is deleted from the sample, this discrepancy is based on the Mahalanobis distance, so the process is robust to changes in scale or position. (See Chapter 2 for details about the discriminator function)

This splitting process is iteratively repeated until the resultant groups are all of sizes smaller than a minimum size. Following the guidelines of Peña et al. (2004), the minimum size is set as $n_0 = p + log(n - p)$ where as usual p is the number of variables and n is the sample size. As a result of the splitting process, we get a set of basic groups, all of them of relatively small size and internally homogeneous.

Given the structure of the basic groups, it is usual that the number of groups is bigger than the actual number of clusters in the data, so a recombination process is needed. We propose the use of the dip test to contrast if two basic groups conform an unimodal sample or not. The idea behind it is that if two basic groups are part of the same original clusters, they should share the same mode.

One of the limitations of the original implementation of the dip test is that is only applicable to univariate samples, so when the dimension of the problem is greater than one, we need to project the data into one dimension before performing the test. For each pair of basic groups, the procedure tests if they are unimodal (and they should be merged), or not. To do so, the natural election for the projection is the Fisher's linear discriminator direction, since it maximize the separation of the groups to be tested. In this case two groups should be merged if even in the projection which separate them the most, they still show one mode (See Section 3.4 for a discussion about the choose of a good direction for the projection).

The output of the test is the value of the dip statistic and the associated p-value calculated with the simulation performed by Maechler (2013), who corrected the original code proposed by Hartigan (1985). The quantiles were obtained using 1000001 samples for each n, and a summary of they are shown in table 3.1.

Given that all possible combinations of basic groups have been tested via the dip statistic, we propose the use of a graphical tool to identify if the groups should be merged or not. To do so, we plot all groups as nodes in a network, and when for a pair of groups the null hypothesis of unimodality is not rejected (i.e. the groups can be merged) the two nodes will be connected by a line. Varying the minimum level of significance α over the set $[0; 1]$ is possible to see the evolution of the grouping process, although the usual $\alpha = 0.1, 0.05$ and 0.01 should unveil the structure of the data set.

After combined, the remaining observations which were not previously assigned to the basic sets, can be incorporated to the resulting sets by using the criteria of smaller Mahalanobis distance.

TABLE 3.1: Table of quantiles from a large simulation for Hartigan's dip test

	0.01	0.05	0.1	0.5	0.9	0.95	0.99	0.995	0.999
4	0.1250	0.1250	0.1250	0.1250	0.1874	0.2073	0.2318	0.2373	0.2444
5	0.1000	0.1000	0.1000	0.1216	0.1768	0.1864	0.1965	0.1982	0.1996
6	0.0833	0.0833	0.0833	0.1231	0.1591	0.1648	0.1919	0.2021	0.2195
7	0.0714	0.0726	0.0817	0.1178	0.1442	0.1599	0.1841	0.1910	0.2023
8	0.0625	0.0739	0.0820	0.1110	0.1418	0.1540	0.1730	0.1790	0.1945
9	0.0613	0.0733	0.0804	0.1042	0.1364	0.1466	0.1642	0.1728	0.1887
10	0.0610	0.0718	0.0780	0.0978	0.1305	0.1396	0.1597	0.1672	0.1806
15	0.0546	0.0610	0.0643	0.0836	0.1101	0.1188	0.1360	0.1425	0.1555
20	0.0474	0.0527	0.0568	0.0733	0.0972	0.1051	0.1206	0.1266	0.1386
30	0.0396	0.0444	0.0474	0.0615	0.0815	0.0882	0.1015	0.1065	0.1172
50	0.0314	0.0353	0.0377	0.0489	0.0649	0.0703	0.0812	0.0853	0.0941
100	0.0228	0.0257	0.0274	0.0355	0.0472	0.0511	0.0590	0.0620	0.0684
200	0.0165	0.0185	0.0198	0.0256	0.0340	0.0368	0.0427	0.0450	0.0497
500	0.0106	0.0119	0.0127	0.0165	0.0219	0.0237	0.0275	0.0289	0.0320
1000	0.0076	0.0085	0.0091	0.0117	0.0156	0.0169	0.0196	0.0206	0.0229
2000	0.0054	0.0061	0.0065	0.0084	0.0111	0.0120	0.0140	0.0147	0.0163
5000	0.0034	0.0039	0.0041	0.0053	0.0071	0.0077	0.0089	0.0093	0.0103
10000	0.0024	0.0027	0.0029	0.0038	0.0050	0.0054	0.0063	0.0066	0.0073
20000	0.0017	0.0019	0.0021	0.0027	0.0035	0.0038	0.0045	0.0047	0.0052
40000	0.0012	0.0014	0.0015	0.0019	0.0025	0.0027	0.0032	0.0033	0.0037
72000	0.0009	0.0010	0.0011	0.0014	0.0019	0.0020	0.0024	0.0025	0.0028

3.3 Results

To illustrate the behaviour of the procedure, remember the Old Faithful geyser data set from Figure 1.1, where we can clearly observe two well differentiated groups. If we apply the splitting step we obtain 12 sets and some isolated observations. These basic groups are shown in the Figure 3.1

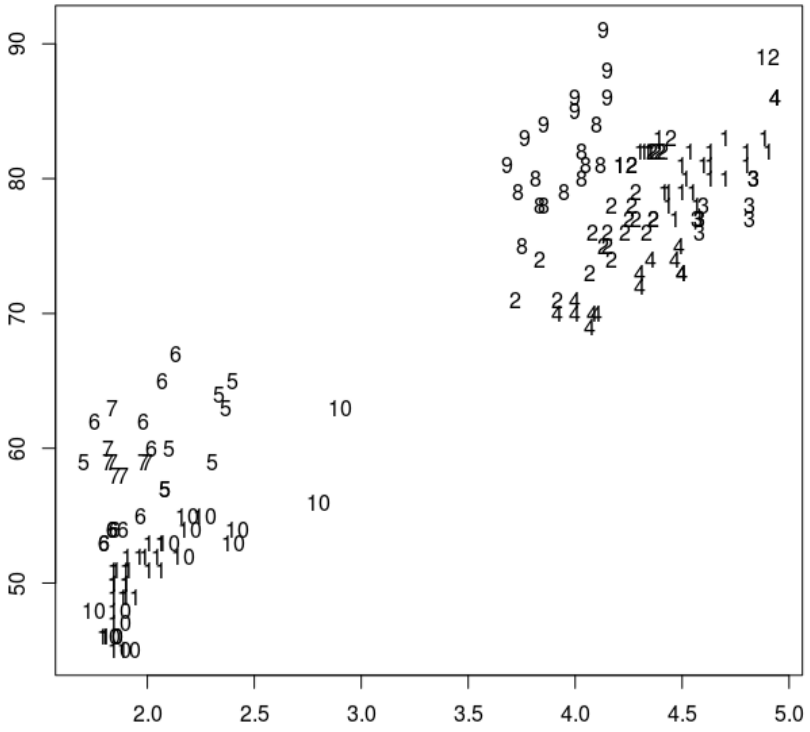

FIGURE 3.1: Basic groups from the Old Faithful data set

The following step is to calculate the dip statistic and the correspondent p-value for each of the $\frac{12 \times (12-1)}{2} = 66$ possible pair of groups. These results are given in table 3.2

TABLE 3.2: Pairwise dip testing of the 12 basic groups obtained from the Old Faithful data set

	Group1	Group2	dip	p-value		Group1	Group2	dip	p-value
1	1	2	0.0413	0.9239	34	4	8	0.1335	0.0002
2	1	3	0.0518	0.8153	35	4	9	0.1214	0.0040
3	1	4	0.0912	0.0169	36	4	10	0.1403	0.0000
4	1	5	0.1091	0.0055	37	4	11	0.1735	0.0000
5	1	6	0.1631	0.0000	38	4	12	0.1192	0.0050
6	1	7	0.1332	0.0001	39	5	6	0.0763	0.4316
7	1	8	0.1002	0.0098	40	5	7	0.0817	0.4889
8	1	9	0.0980	0.0231	41	5	8	0.1461	0.0005
9	1	10	0.1434	0.0000	42	5	9	0.1683	0.0001
10	1	11	0.1522	0.0000	43	5	10	0.0627	0.6105
11	1	12	0.1032	0.0121	44	5	11	0.0825	0.3744
12	2	3	0.0722	0.2995	45	5	12	0.1887	0.0000
13	2	4	0.0541	0.6550	46	6	7	0.0754	0.4519
14	2	5	0.1133	0.0049	47	6	8	0.1916	0.0000
15	2	6	0.1682	0.0000	48	6	9	0.1545	0.0001
16	2	7	0.1328	0.0003	49	6	10	0.0856	0.0787
17	2	8	0.0774	0.1687	50	6	11	0.0808	0.2675
18	2	9	0.1078	0.0099	51	6	12	0.1617	0.0000
19	2	10	0.1299	0.0000	52	7	8	0.1900	0.0000
20	2	11	0.1602	0.0000	53	7	9	0.1954	0.0000
21	2	12	0.1017	0.0212	54	7	10	0.0907	0.0822
22	3	4	0.0647	0.5864	55	7	11	0.1383	0.0024
23	3	5	0.1847	0.0000	56	7	12	0.2023	0.0000
24	3	6	0.1931	0.0000	57	8	9	0.1009	0.0906
25	3	7	0.2207	0.0000	58	8	10	0.1331	0.0001
26	3	8	0.1585	0.0000	59	8	11	0.2121	0.0000
27	3	9	0.1742	0.0000	60	8	12	0.0987	0.1077
28	3	10	0.1399	0.0001	61	9	10	0.1082	0.0123
29	3	11	0.2140	0.0000	62	9	11	0.1692	0.0000
30	3	12	0.1717	0.0000	63	9	12	0.1526	0.0009
31	4	5	0.1249	0.0024	64	10	11	0.0627	0.5464
32	4	6	0.1775	0.0000	65	10	12	0.1249	0.0013
33	4	7	0.1397	0.0003	66	11	12	0.1801	0.0000

If we take a look into two basic sets which belongs to the same cluster, for example, sets 1 and 2, the density plot of their projection into the Fisher direction does not show a bimodal evidence (See Figure 3.2), and the p-value from table 3.2 is 0.9239. In the case of basic groups 2 and 10, the associated p-value is equal to 0, and the corresponding density plot clearly shows two modes. (See Figure 3.3)

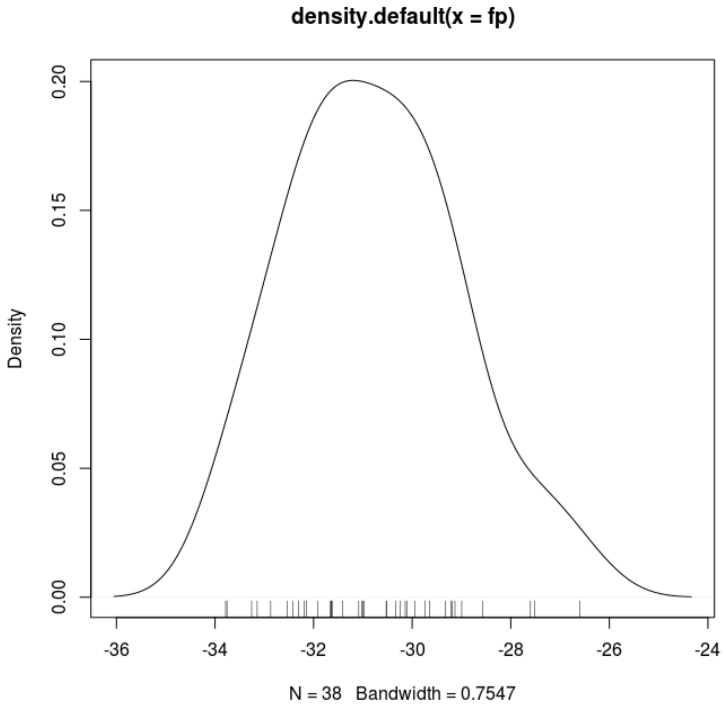

FIGURE 3.2: Density function of univariate projection of basic sets 1 and 2

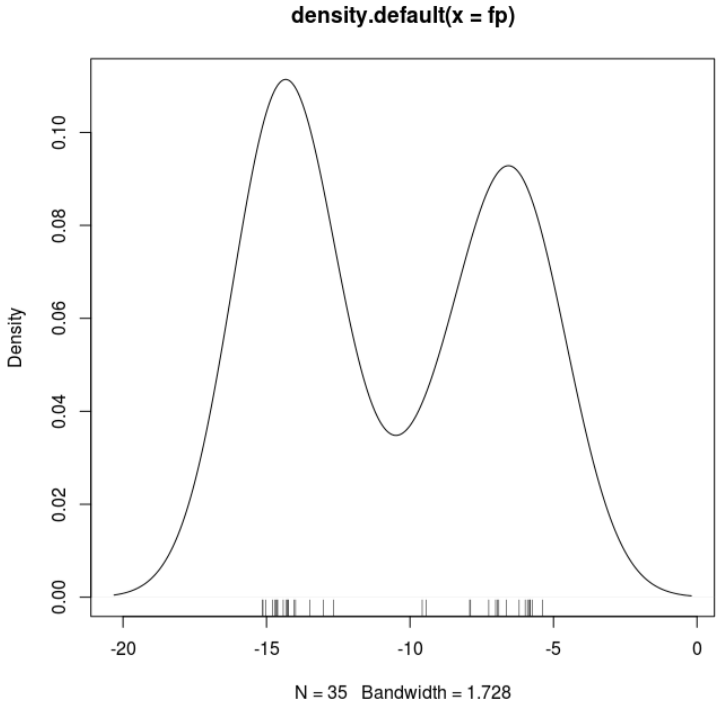

FIGURE 3.3: Density function of univariate projection of basic sets 2 and 10

Graphically, the interaction between all basic sets is shown in the Figure 3.4, where we observed two clearly differentiated groups, one formed by groups 5,6,7,10 and 11; and other by the remaining basic sets, corresponding with the original configuration of the data.

This graphical tool as an exploratory approach, allows also to see different strengths within the groups. For example, the group formed by sets 5-6-7-10-11 seems to be more internally connected than the group composed by basic sets 1-2-3-4-8-9-10, which can be separated into a "strong group" of sets "1-2-3-4" and other formed by 8-9-12.

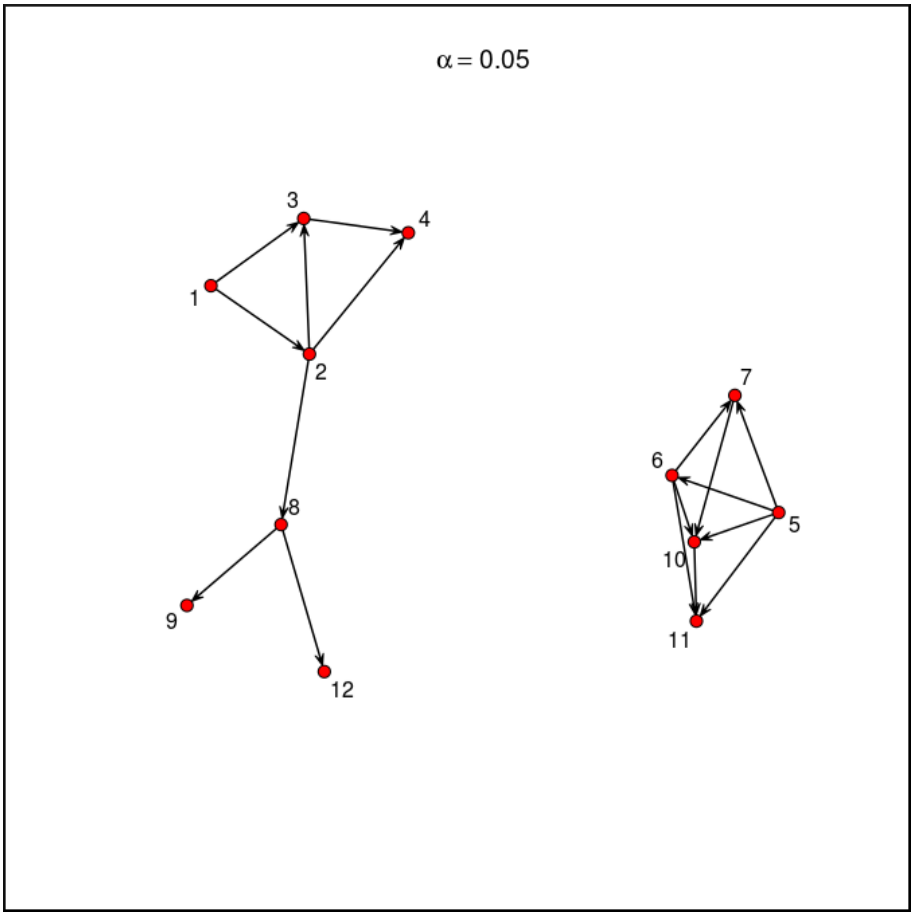

FIGURE 3.4: Dip test network for the Old Faithful data set, $\alpha = 0.05$

As a second example, we consider a case when the data set is not linearly separable. In Figure 3.5 we show the simulation of two half-moons, each of them consisting of 250 data points in two dimensions. After the splitting procedure, we find 19 basic groups (See Figure 3.6), while the graphical results of the dip test are shown in the Figure 3.7.

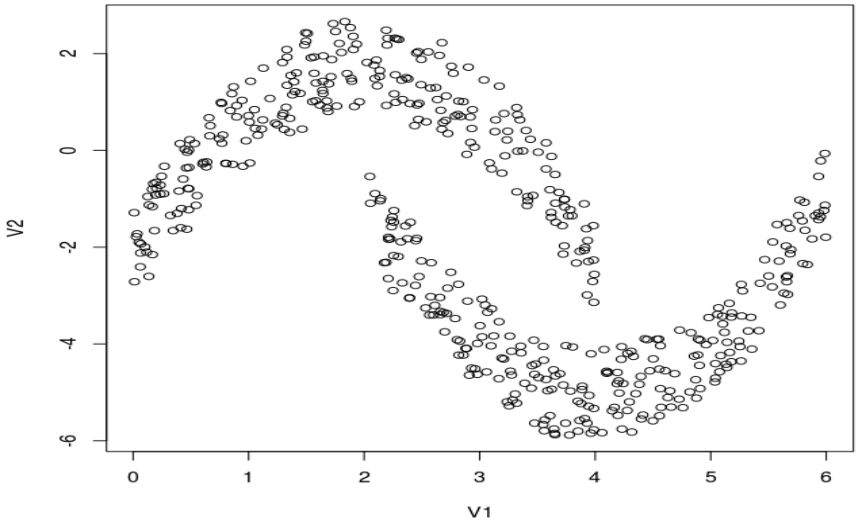

FIGURE 3.5: The two half moons data set

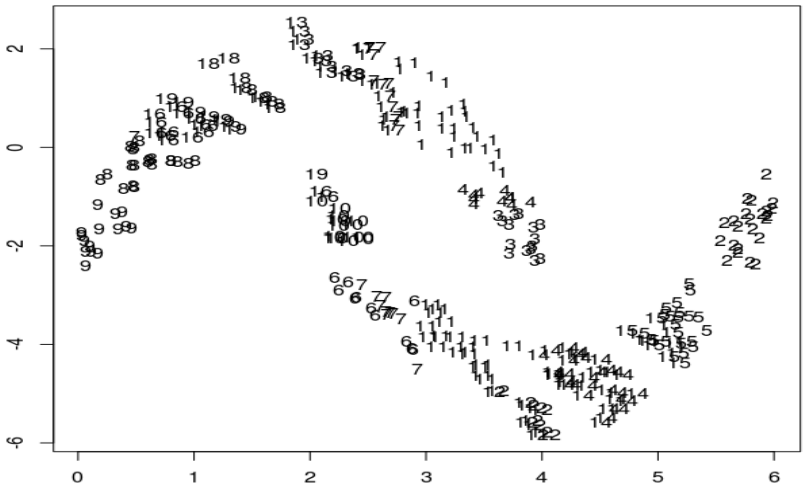

FIGURE 3.6: Basic groups of the two half moons data set

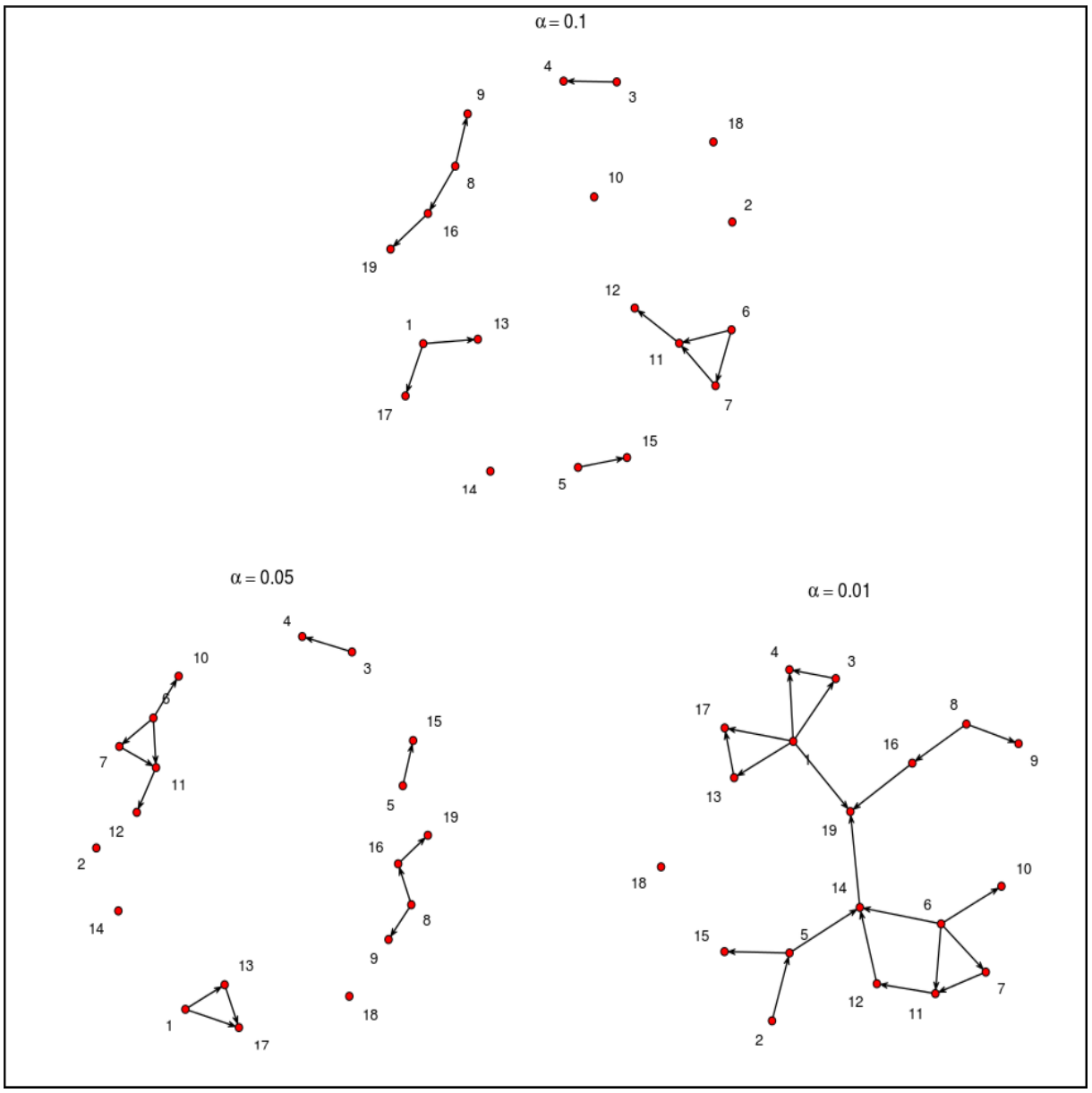

FIGURE 3.7: Dip network for the two half moons data set for $\alpha = 0.1, 0.05$ and 0.01

When $\alpha = 0.1$, the procedure detects 5 combination of groups and other 4 single groups, not detecting yet the structure of two groups in the data. A similar situation occurs when α is decreased to 0.05, but when we consider $\alpha = 0.01$ the two clusters appears, one in the upper half of the network formed by groups 1-3-4-8-9-13-16-17-19 and other in the lower half composed of groups 2-5-6-7-10-11-12-14-15, plus an isolated group (18) unveiling the more complex structure of the data. Notice that these two clusters are connected by the group 19, reflecting a problem

in the partition process, because that group incorrectly includes observations from the two half moons (See Figure 3.6).

In complex data sets, the partition process need to be adapted, for example including a cleaning process as we will discuss in next Chapter, therefore, smaller values of α are needed in order to properly reflect the original clusters of the data. This procedure should be intended as exploratory, where is possible to observe the evolution of the combining process while varying the confidence levels.

3.4 Discussion

In order to recombine multivariate subpartitions based on unimodality tests there are two main approaches: First, to keep the dimension of the problem and to look for an appropriate multivariate modality detection, or second, to use a simpler univariate modality test but to choose a good projection direction reducing the dimensionality of the data. We will briefly discuss this two alternatives and justify the election we made for this research.

3.4.1 Multivariate modality tests

Besides the dip test, Hartigan tried to extend its work to a multivariate framework. On his publications we can find three proposals in that direction, the tests "span" (Hartigan 1988), "RUNT" (Hartigan and Mohanty 1992), and "MAP" (Rozál and Hartigan 1994). All of them are based on a hierarchy of similarities: starting from the n classes corresponding to the n initial points of the sample, and finishing with all data points in one class, the distance between two classes A,B is defined as the smallest distance between an observation from the class A and another observation from the class B.

The RUNT test is based on the fact that for a bimodal distribution is expected that the two modes of the distribution are merged in the last step of the hierarchy, while the span test is a generalization of the dip test where the empirical function

$F_n(x)$ is the proportion of points x_i such that $x_i \preceq x$. Starting from a random root point $r = x_k$, $x \preceq y$ if x is further away from r in the hierarchy. Finally, the MAP test is based on the Minimal Ascending Path, calculated from a MAP Spanning Tree which is a tree such that the length of the links are non-increasing from any link to a root node.

Departing from hierarchy trees, another more recent researches have been focused in the mode detection problem:

Burman and Polonik (2009) assume the data is coming from an unknown distribution with isolated modes. The idea of the method is first, to find potential mode candidates and second, determine if they represent different modal regions via pairwise statistical tests. A modal region is defined as a set R_y with $y \in R_y$, and $f(y + \alpha x)$, with $\alpha \in [0, 1]$, decreasing $\forall x \in Ry$, being y a mode of f.

The first candidate W_1 to be a mode is selected as the observation which have its k_1 neighbour closer. Formally, if $\hat{d_n}(x_j)$ is the distance between an observation $x_j, j = 1, ..n$ and its k_1 nearest neighbour, then:

$$W_1 = \underset{x_j}{arg\ min}\ \hat{d}_n(x_j) \tag{3.2}$$

The second candidate is obtained in a similar way but deleting from the sample the previous candidate and its k_2 neighbours, and the procedure is continued until no more candidates are found.

As a second step, the list of candidates is purged, keeping only those observations which does not significantly differ from the mean of its k_2 neighbours, using a Hotelling's test and assuming multivariate normality.

Finally, the candidates are pairwise tested to belong to the same modal region, by considering the existence of "antimodes" between them. One of the possible tests the authors propose to compare two candidates x and y is the following statistic:

$$\hat{SB}(\alpha) = p\left[\log \hat{d_n}(x_\alpha) - \max\left\{\log \hat{d_n}(x), \log \hat{d_n}(y)\right\}\right] \tag{3.3}$$

where $x_\alpha = \alpha x + (1-\alpha)y, 0 \leqslant \alpha \leqslant 1$. The authors propose to reject the null hypothesis when $\hat{SB}(\alpha) \geqslant \sqrt{\frac{2}{k_1}}\Phi^{-1}(0.95)$, being Φ the c.d.f. of the multivariate normal distribution

Einbeck (2011) develops a technique for multivariate mode detection, although the main objective of their research is focus on a cluster analysis algorithm. The base of the mode detection is the work of Cheng (1995), who defined the "mean shift" as the shift necessary to move a point $x \in \mathbb{R}^p$ towards the local mean around this point.

Let K be a p-variate kernel function (usually Gaussian), and $H = diag(h_1^2, h_2^2, ..., h_p^2)$, with $h_j > 0$ a bandwidth matrix, then:

$$K_H(x) = |H|^{-1/2}K(H^{-1/2}x) \tag{3.4}$$

the local mean and the mean shift are then defined as:

$$\mu_H(x) = \frac{\sum_{i=1}^n K_H(x_i - x)x_i}{\sum_{i=1}^n K_H(x_i - x)} \tag{3.5}$$

$$S_H(X) = \mu_H(x) - x = \frac{\sum_{i=1}^n K_H(x_i - x)(x_i - x)}{\sum_{i=1}^n K_H(x_i - x)} \tag{3.6}$$

For a given distribution function f, and bandwidth H, at a mode m_H of f, $S_H(m_H) = 0$, then $\mu_H(m_H) = m_H$. The authors recall all points satisfying that condition as "Local principal points"

In order to find those local modes, Cheng (1995) proved that the sequence $m_l, l \geq 0$ will converge to a local principal point m_H, with $m_0 = x$, and $m_{l+1} = \mu_H(m_l)$., and this mean shift sequence is iterated the for all data observation.

The application to our original problem is now clear, given two partitions we will recombine them if we found only one mode on its merged set, and keep them separate in other case.

Recalling the Old Faithful example, where the basic groups are plotted in Figure 3.1, we will apply the procedure of Einbeck (2011), since its method is already implemented in an R package (Einbeck and Evers 2012). However, for further research to build our own implementation of Burman and Polonik (2009) seems to be feasible in order to compare multivariate mode detection methods.

Several parameters need to be fixed in the procedure, including *taumin*, *taumax* and *gridsize*, all of them related with the grid of bandwidths where the search for modes is focused. Default options are $taumax = 0.02$, $taumin = 0.5$, and $gridsize = 25$, although its application to the Old Faithful basic groups does not properly recognise the two clusters under this parameters (Figure 3.8). Two other parameter combination are shown in Figures 3.9 and 3.10, being the last one which correctly identify the clusters.

As is shown in the figures, the procedure is highly sensitive to the parameters, and its interpretation is not as clear as the dip test proposed in the previous sections. At the same time the parameters cannot be dynamically adjusted from an "all connected" to a "none connected" framework in a simple way, hindering its visualization. Nevertheless, for higher dimensions this procedure take advantage since the projection can produce high loss of information.

3.4.2 Directions to project the data

Given two multivariate candidate groups for recombining, the choice of the Fisher's direction to project the data in the proposed procedure is natural, since it maximizes the separability of the groups, and has been long used in classical methods as discriminant analysis. For our problem, that choice implies the most conservative scenario, because it tests unimodality even in the case where the separation between groups is maximum.

In the context of cluster analysis, the search for interesting directions to project the data and keep the structure of it has been widely used as a way to avoid the dimensionality curse (Friedman and Tukey 1974; Friedman 1987). The choice of Fisher's direction is also supported by the literature: Peña and Prieto (2001) proposed the direction that minimize the kurtosis as appropriate for cluster analysis, and later, Peña, Prieto, and Viladomat (2010) proved that given the kurtosis matrix, the subspace orthogonal to the eigenspace associated to an eigenvalue with multiplicity $p-k+1$ is Fishers linear discriminant subspace. Similar results can be found in Caussinus and Ruiz-Gazen (1994) and Caussinus and Ruiz-Gazen (1995), where the Fishers subspace is obtained from the k largest eigenvectors of a Generalized Principal Components matrix, or Tyler, Critchley, Dümbgen, and Oja (2009) who proved that it can be generated from eigenvectors of affine equivariant scatter matrices.

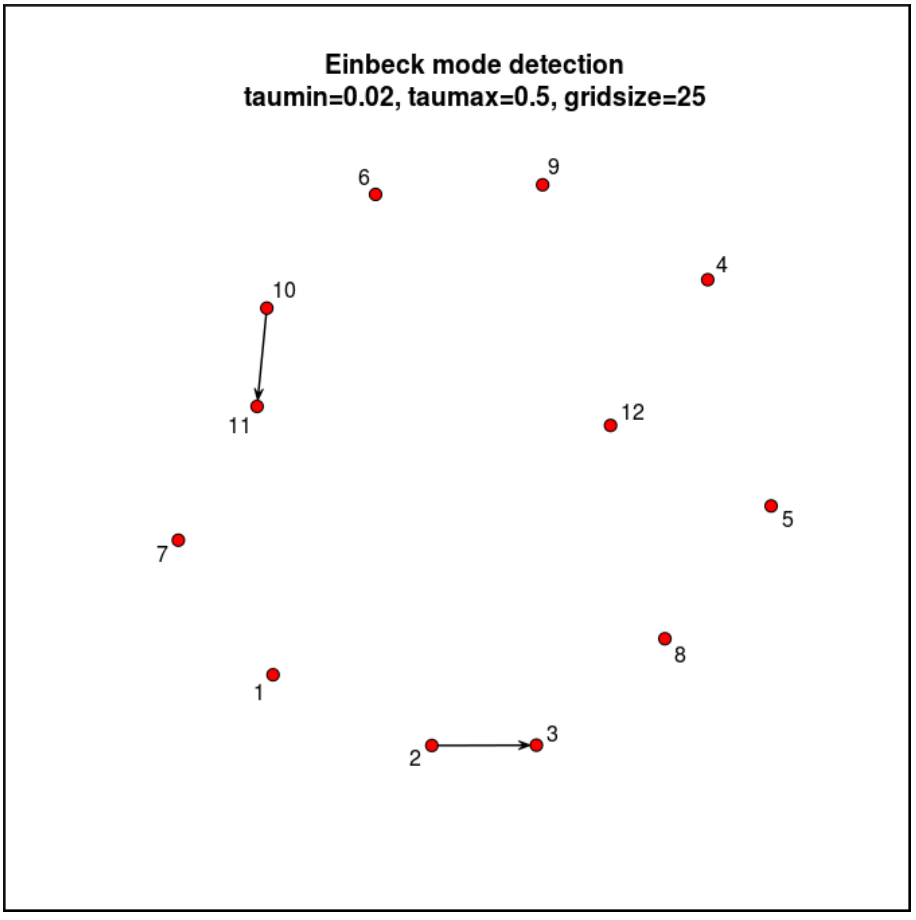

FIGURE 3.8: Einbeck mode detection test with default parameters

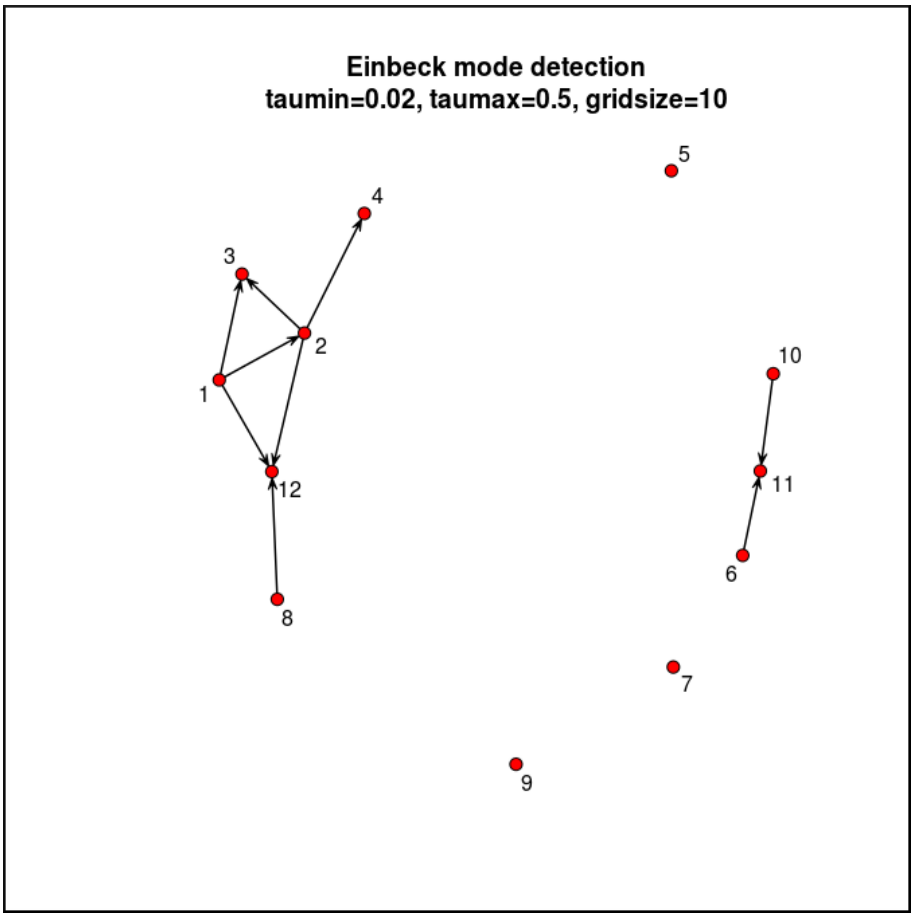

FIGURE 3.9: Einbeck mode detection test, gridsize decreased

The choice of Fisher's direction is optimal when we actually know the two partitions we want to test for recombine. Only under the assumption of no knowledge about the basic groups, one of the alternative directions presented here can be considered, for example in the case of a splitting step, where we can project the data and split into groups until no bimodality can be detected.

3.5 Conclusions

We have developed a method to split a data set using the discriminator function and recombine the obtained groups to find the final configuration incorporating the dip-statistic to test for unimodality. Also, we presented a graphical tool which allows to see the evolution of the merging procedure, and unveil which groups are

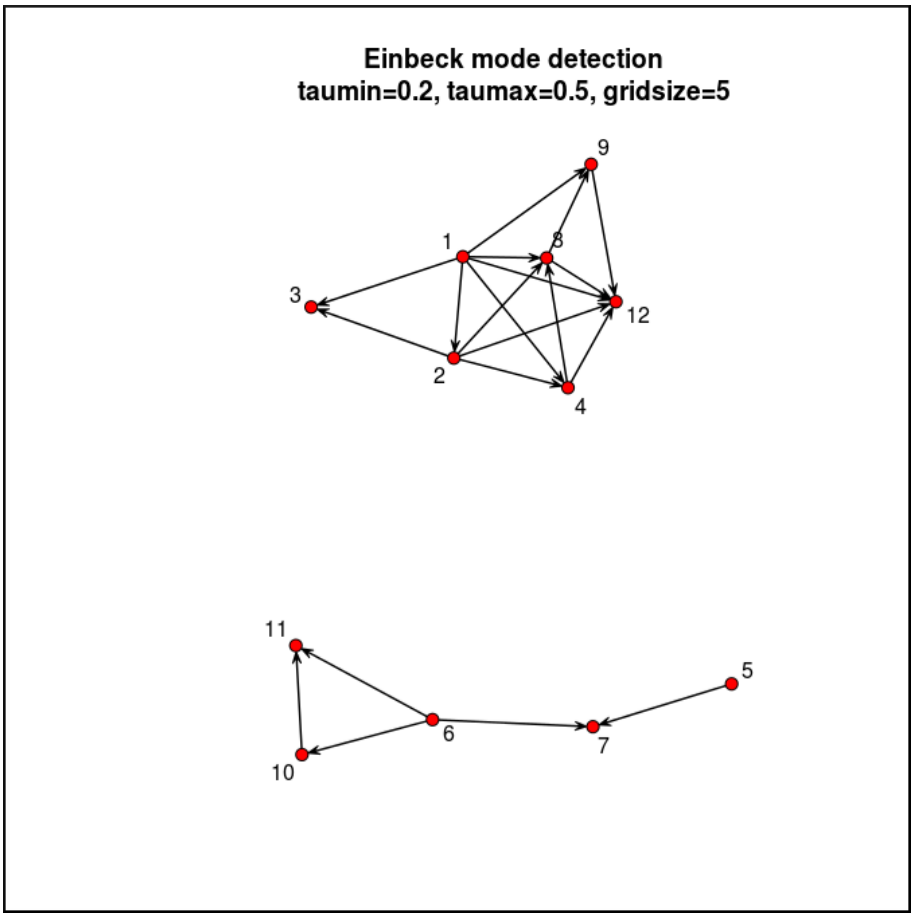

FIGURE 3.10: Einbeck mode detection test, gridsize decreased and taumin augmented

more internally connected. The results show that the proposed technique can be a useful tool for exploratory research, since it allows to dynamically vary the level of significance to visualise the merging behaviour of the procedure.

The method have two issues which must be taken into account when applying it, which are also present in other dip-statistic based approaches: the validity and interpretation of p-value, and the chosen projection technique for multivariate data.

The obtained p-values does not hold the assumption of independence of a standard hypothesis test, because the partitions we test for unimodality are obtained from a previous methodology, and as we saw in Chapter 1, they are dependent in the sense they are disjoint by construction. Therefore, we test the same data several times, because we compare each basic group against all the rest.

Nevertheless, even without the traditional interpretation of p-values, they can be used to show the behaviour of the merging when we modify the minimum level α from 1, where no groups are connected, to 0, where there is a connection between all groups. The most similar groups will merge in values close to 1, and clearly disjoint groups will not merge until values below 0.01.

In the other hand, it is important to notice that some useful information of the structure of the real data can be lost in the reduction of dimensionality. This is specially relevant in complex data sets or high dimension problems, and in this context, multivariate mode detection techniques, as those we reviewed in the discussion section, should be preferred.

Chapter 4

Recombining partitions from multivariate data: A clustering method based on Bayes factors

4.1 Introduction

This chapter deals with the problem of recombining partitions in multivariate data. In this context, we present a new clustering methodology which, based in a strategy of splitting, cleaning and recombining, is able to detect groups inside a data sample. As splitting rule we use the discriminator function, where points sharing the same discriminator are classified into the same group, defining in this way a partition of the sample. For cleaning we detect and purge the outliers of each group, and finally for recombining, we propose the use of a Bayes factor to weight the likelihood of the sample given two models: one where all data is generated from a single distribution, and other when the distribution is a mixture estimated from the obtained partition.

We follow the same split and recombine approach as the SAR algorithm for exploratory data analysis proposed by Peña et al. (2004), reviewed in Section 1.4.1, which split the sample using an heterogeneity measure based in the Mahalanobis

distance, to later enlarge the resulting small groups one data point at a time to form several possible data configurations. Nevertheless, we modify the splitting and recombining processes from the SAR algorithm, incorporating to the splitting an outlier detection process which tries to avoid mixing observations from different clusters in the same basic group. Second, in the recombining process we merge the groups obtained in the splitting, while the SAR test each observation to be incorporated to a group. In this way, we use the information given by those original partitions, increasing the efficiency of the procedure, and obtaining only one data configuration as an output.

4.1.1 Brief literature review

The split and recombine methodology has been followed by several authors in cluster analysis. In fact, classical methods as k-means proposed by MacQueen (1967) can be considered as a "split and recombine" method, although the split process is based only in a few observations that are considered the starting points for the aggregation. A general review of k-means and other classical cluster methods in the context of this research was presented in Section 1.2.

Some algorithms take as input the partition process from outside procedures as those which are focused on recombining normal samples, particularly useful when a mixture of normal distributions, as those obtained by the M-clust algorithm (Fraley and Raftery (2002)), overestimates the real number of clusters in a sample. For example, Tantrum et al. (2003) propose the use of the dip test of unimodality (Hartigan and Hartigan 1985) to recombine such mixtures (See Chapter 3), and Baudry, Raftery, Celeux, Lo, and Gottardo (2010) propose the use of the Integrated Completed Likelihood (ICL) criteria, established by Biernacki, Celeux, and Govaert (2000) where is assumed than a non-observable component containing the assigning labels of the data to the groups can be incorporated to the likelihood. This criteria penalize the BIC by the Mean Entropy leading to a smaller number of components than BIC.

Beyond M-clust, a set of methods to merge Gaussian distributions based on misclassification probabilities are proposed by Hennig (2010a). The first is based on the Bhattacharaya distance defined by Fukunaga (1990), which measures the Bayes misclassification probability between two distributions. The second is called "DEMP method" and uses directly estimated misclassification probabilities, being these probabilities given by the EM algorithm. The last one is the "prediction strength method" where the misclassification is calculated splitting the data in two halves and use one half to predict the cluster membership of the second half.

More recent approaches to the problem of merging Gaussian components can be found in literature, including a topology based methodology using manifolds (Hennig 2010b), averaging the clustering results of several models (Wei and McNicholas 2012), applying k-means over the components means (Li 2005), or measuring the Kullback-Leibler divergence between two distributions (Popović, Janev, Pekar, Jakovljević, Gnjatović, Sečujski, and Delić 2012). A deeper review of this topic can be found in Hennig (2010a).

Atkinson and Riani (2007) follows a similar approach than Peña et al. (2004) in their clustering proposal. The authors use a forward search starting from random subsets of the data sample and calculate robust Mahalanobis distances between each element and the initial subset. If the subset is of size m, in each step the starting group is enlarged in one element by selecting the m+1 smaller distances, recalculating again the distances until all data sample is included. By plotting the Mahalanobis distances is possible to identify the original groups of the sample observe the peaks produced in the distances when observations belonging to different clusters are added to the subset.

A more recent clustering methodology using a split and recombine approach can be found in Fraiman, Ghattas, and Svarc (2011) who propose an algorithm inspired in the CART technique for supervised classification problems (Classification and Regression Trees, Breiman, Friedman, Olshen, and Stone 1984). The process is done defining a nodes structure starting with one node with all data set and successively splitting the space where the data set lays, perpendicularly to the axes

of each dimension of the data set, conforming a binary tree. In a second stage, a merging process is done via combining the different nodes based on distances between each pair of nodes, and setting an expected number of clusters, or a cut-off as a stop rule.

Casella and Fuentes (2009) propose a different approach to detect clusters inside a data sample. They establish a test for the hypothesis $H_0 : \kappa = 1$ vs. $H_1 : \kappa = k$, where κ represents the number of clusters. Our procedure is in the same direction, so we will review it in deep.

Let $X = X_1, X_2, ..., X_n$ be the data sample, where each $X_i, i = 1, ..., n$ is an element of p-dimensions. Then a partition ω_k is a n-dimension vector which assigns each element of the sample X to one of the k groups, representing a way to cluster n elements into k groups.

For example, when $n = 3$, we have $X = X_1, X_2, X_3$ and the set of possible partitions are for $k = 1 : \{(X_1, X_2, X_3)\}$, leading to $\omega_1 = \{(1, 1, 1)\}$ for $k = 2 : \{(X_1, X_2), (X_3)\}, \{(X_1, X_3), (X_2)\}, \{(X_1), (X_2, X_3)\}$, so $\omega_2 \in \{(1, 1, 2), (1, 2, 1), (1, 2, 2)\}$ and for $k = 3 : \{(X_1), (X_2), (X_3)\}$, there is only one possible partition $\omega_3 = \{(1, 2, 3)\}$.

The number of ways to divide a set of n objects into k non-empty subsets, is called the "Stirling number of the second kind", and can be calculated as:

$$S_{n,k} = \frac{1}{k!} \sum_{j=0}^{k-1} (-1)^j \binom{k}{j} (k - j)^n. \tag{4.1}$$

From Equation (4.1) is clear that the Stirling number of the second kind grows exponentially, even with relatively small sizes of n and k. For example $S_{20,3} = 580,606,446$. For more details, a complete review of this measure can be found in Moll (2012).

Casella and Fuentes test is based in a Bayes factor associated with the null and alternative hypotheses as given by Equation (4.2), where $m(X|\kappa = k)$ represents the marginal of the likelihood of the data, X, given that there are k clusters.

$$BF = \frac{m(X|\kappa = k)}{m(X|\kappa = 1)} \tag{4.2}$$

Considering the total number of all partitions, given by $S_{n,k}$, the Bayes factor can be written as:

$$BF = \sum_{\omega \in S_{n,k}} \frac{m(X|\omega)}{m(X|\omega_1)} \frac{\pi(\omega)}{\pi(\omega_1)} \tag{4.3}$$

where $\pi(\omega)$ is the prior probability of the partition ω, and the posterior probability of H_0 is calculated then in terms of the Bayes factor, $P(H_0|X) = 1/(1 + BF)$.

For each $\omega \in S_{n,k}$ the authors assume that the observations in the cluster j are distributed $N(\mu_j, \Sigma_j)$, so the likelihood of the sample and the marginal given a partition ω can be described by Equations (4.4) and (4.5) respectively.

$$L(\mu, \Sigma, \omega|X_1, X_2, \ldots, X_n) = \prod_{j=1}^{k} \prod_{l=1}^{n_j} N(X_l{}^{(j)}|\mu_j, \Sigma_j) \tag{4.4}$$

$$m(X|\omega) = \iint \prod_{j=1}^{k} \prod_{l=1}^{n_j} N(X_l{}^{(j)}|\mu_j, \Sigma_j) \cdot p(\mu_j, \Sigma_j) d\mu_j d\Sigma_j \tag{4.5}$$

Finally, as priori distribution for the mean and variance, the authors propose the use of: $p(\mu_j, \Sigma_j) = p(\mu_j|\Sigma_j) \cdot p(\Sigma_j)$ with:

$$p(\mu_j|\Sigma_j) \sim N(\mu_0^{(j)}, \tau^2 \Sigma_j) \tag{4.6}$$

where $\Sigma_j = diag(\sigma_{1j}^2, \sigma_{2j}^2 \ldots, \sigma_{rj}^2)$ and $\sigma_{rj}^2 \sim InverseGamma(a, b)$ with fixed values for $a = 2.01$ and $b = (a-1)^{-1}$

One of the issues of this methodology is that the sum of all possible partitions is in general too big, so is necessary to apply a Metropolis Hasting algorithm to sum over the subset of partitions that contribute more to the total sum in Equation (4.3). Nevertheless, the main idea of comparing models by a Bayes factor is an

useful tool to decide whether to recombine a set of groups to detect clusters, and we will follow this approach in the next sections.

4.1.2 Structure of the chapter

The chapter is structured as follows: in Section 4.2 we will develop the fundamentals of the recombining method, where we propose the use of a Bayes factor to compare two possible models explaining the data; In Section 4.3 we describe the algorithm which integrate the splitting and recombining processes; in section 4.4 we show the results of the application of the proposed method to four different data configurations; and in section 4.5. we present some conclusions.

and finally we show the results of the application of the proposed method to four different data configurations and conclusions in Sections 4.4 and 4.5 respectively.

4.2 The splitting, cleaning, and recombining proposals

As in the original SAR process, the core of our methodology is based in the use of splitting and recombining procedures, plus a cleaning step between them. The splitting process will be based on the discriminator function, the cleaning in an outlier detection process based on robust Mahalanobis distances, and finally for recombining we propose the use of Bayes factors. In this section we will review these procedures.

4.2.1 Splitting

The basic measure for the splitting is the discriminator function, already introduced in Subsection 1.4.2, where two observations y_i and y_j are assigned to the same group if they share a discriminator i.e. $y_l(y_i) = y_l(y_j)$, where y_l is obtained

from Formula (1.4). This measure allows to identify similar observations since as highlighted in Peña et al. (2004), "if two observations are identical they will have the same discriminator and if they are close they also will have the same discriminator". Since the splitting process is virtually identical to the equivalent in the SAR algorithm, we will not deepen on the details of the discriminator function that can be found in the above references and in Rodriguez (2002).

4.2.2 Cleaning

The outlier detection problem has been widely studied by the literature and is always a hot topic in statistical applications, because, as we briefly discussed in Chapter 1, the presence of outliers can bring inference complications as biased estimation, loss of efficiency, or bad predictions. Recent applications of outlier detection can be found in different areas such as cancer diagnostic (Kothari, Wei, and Shankar 2013), climate change (Cho, Oh, Kim, and Shim 2013) or wireless networks (Branch, Giannella, Szymanski, Wolff, and Kargupta 2012).

To develop a new efficient method of outlier detection is far beyond the goals of this thesis, and even a comprehensive literature review of this topic can take hundreds of journal articles and books. Good recent examples of such reviews are Pahuja and Yadav (2013), Hodge and Austin (2004) and the book of Aggarwal (2013).

Among this big number of possibilities of outlier detection methodologies that we could incorporate to our proposal, the first and natural option to consider was the original SAR outlier detection presented in Peña and Tiao (2006) . In the original SAR process, the outlier detection step (See Section 1.4.1 for details) is performed before the splitting using the measures defined by Equations (1.1) and (1.3), based on Mahalanobis distances.

A more classic approach to the problem of outlier detection is the work of Rousseeuw (1985) who also propose the use of Mahalanobis distances to detect outliers. Under (multivariate) normality, the Mahalanobis distances are approximately distributed

as a chi-square with p degrees of freedom (χ^2_p), but given that outliers can influence in those distances, is necessary to estimate them using a robust procedure. The method, known as MCD (Minimum covariance determinant) estimate the covariance matrix by the subset of h observations which minimises its determinant. Rousseeuw and Driessen (1999) shows that the MCD method is a computationally fast algorithm that can be used to calculate robust Mahalanobis distances based on those estimators and detect outliers. In practice, we use the implementation of the method proposed by Filzmoser, Garret, and Reimann (2005) who also incorporate flexible critical values for those robust distances and that is available under the "mvoutlier" library for the R statistical language (Filzmoser and Gschwandtner 2013).

Let $G_n(u)$ be the empirical distribution of the squared robust distances RD_i^2 calculated with the MCD method, is possible to compare the tails of that distribution and the theoretical distribution χ^2_p to detect outliers. The tails will be defined by $\delta = \chi^2_{p;1-\alpha}$ for a certain small α, so the departure of the empirical from the theoretical distribution in the tails is given by:

$$p_n(\delta) = \sup_{u \geq \delta} (\chi^2_p - G_n(u))^+ \tag{4.7}$$

where $+$ indicates only the positive differences. Using this measure is still important to distinguish between extremes of the distribution and real outliers. To do so, a critical value (p_{crit}) is proposed by Filzmoser et al. (2005):

$$p_{crit} = \frac{0.24 - 0.003p}{\sqrt{n}} \text{ for } p \leq 10. \tag{4.8}$$

$$p_{crit} = \frac{0.252 - 0.0018p}{\sqrt{n}} \text{ for } p > 10. \tag{4.9}$$

Finally, the threshold value for the outliers is determined by

$$c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta)) \tag{4.10}$$

where

$$\alpha_n(\delta) = \begin{cases} 0, & \text{if } p_n(\delta) \leq p_{crit}(\delta, n, p). \\ p_n(\delta), & \text{if } p_n(\delta) > p_{crit}(\delta, n, p). \end{cases} \tag{4.11}$$

To depart from multivariate normality assumptions, a third option could be a more general algorithm of outlier detection in multivariate analysis proposed by Peña and Prieto (2001). The proposal is based on the idea of using projections to identify outliers, where each outlier must be an extreme point along the direction from the mean of the uncontaminated data to the outlier. In order to determine the direction of the projections, the authors claim that the presence of outliers in the projected data will imply particularly large (or small) values for the kurtosis coefficients, so they propose to use those directions that maximize or minimize the kurtosis.

4.2.3 Recombining

Given a partition $\omega_k = (l_1, l_2, ..., l_n)$ where $l_i \in \{1, 2, ...k\}, i = 1, 2, ...n$ are the labels assigning n data points $X_1, X_2, ..., X_n$ into $k > 1$ clusters, generated by the splitting process, we use a Bayes factor to compare the probability of the observed data given that partition against the data given the partition $\omega_1 = (1, 1, 1, ..., 1)$, implying all data points come from the same cluster in a similar way as used by Casella and Fuentes (2009).

Under the framework of recombine cluster subpartitions (or basic groups) as those obtained by a splitting procedure, we improve the Casella and Fuentes' Bayes factor in two ways:

a) We do not need to sum over all possible partitions or perform an importance sampling to estimate the Bayes factor, since we can use the information obtained by the splitting process. For example, consider two basic groups of sizes n_1 and

n_2, so the Bayes factor to test if these two basic groups should be recombined will be:

$$BF = \frac{m(X|\omega_2) \cdot \pi(\omega_2)}{m(X|\omega_1) \cdot \pi(\omega_1)} \tag{4.12}$$

, where $\omega_2 = (\underbrace{1, 1, 1, ..., 1}_{n_1}, \underbrace{2, 2, 2, ..., 2}_{n_2})$ and $\omega_1 = (\underbrace{1, 1, 1, ..., 1}_{n_1+n_2})$

This approach has two main advantages: is more efficient in terms of computation time, and it also uses the information obtained by the partition process, which can be relevant to find the underlying structure of the data.

b) As a prior distribution for the mean and variance to derive the marginal given a certain partition, Casella and Fuentes use a restrictive approach where the covariance matrix is assumed to be diagonal. We propose to use a more flexible but also simple alternative, the use of the standard non informative given by Jeffreys:

$$p(\mu_j, \Sigma_j) \propto |\Sigma|^{\frac{-p-1}{2}}$$

Under this priori, Geisser (1964) shows that the posterior probability of an observation given a cluster defined by a $N(\bar{x}_i, S_i)$ is:

$$\begin{aligned} p(z|\bar{x}_i, S_i) &= \iint p(z|\mu_j, \Sigma_j) \cdot p(\mu_j, \Sigma_j|\bar{x}_i, S_i) d\mu_j d\Sigma_j \\ &= \left\{ \frac{N_i}{N_i + 1} \right\}^{p/2} \frac{\Gamma\{\frac{1}{2}(N_i)\}}{\Gamma\{\frac{1}{2}(N_i - p)\}|(N_i - 1)S_i|^{1/2}} \\ &\times \left[1 + \frac{N_i(\bar{x}_i - z)' S_i^{-1}(\bar{x}_i - z)}{(N_i + 1)(N_i - 1)} \right]^{-1/2(N_i)} \end{aligned} \tag{4.13}$$

Given a partition ω_k defined by k subsamples from the splitting process, in order to test if we will combine them ($H_0 : \kappa = 1$ *vs* $H_1 : \kappa = k$,), we will use the Bayes factor given by Equation (4.12), where $m(X|\omega)$ is the likelihood of the data

given by partition ω, under the assumption that each of the groups in the partition follows a multivariate normal distribution, and using non informative priors for μ and Σ as given by Equation (4.13).

$$m(X|\omega) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\left\{\frac{N_i}{N_i+1}\right\}^{\frac{p}{2}} \frac{\Gamma\{\frac{1}{2}(N_i)\}}{\Gamma\left\{\frac{(N_i-p)}{2}\right\}|(N_i-1)S_i|^{\frac{1}{2}}} \times\left[1+\frac{N_i(\bar{x}_i-y_{ij})'S_i^{-1}(\bar{x}_i-y_{ij})}{(N_i+1)(N_i-1)}\right]^{\frac{-N_i}{2}} \tag{4.14}$$

In a similar way to Casella and Fuentes (2009), we use the marginal distribution of the number of clusters in a Dirichlet process proposed by Pitman (1996) as priors for partitions $\pi(\omega)$. In this configuration, the priors only depend on the number of elements in each group of the partition.

$$\pi(\omega_k) = \frac{\prod_{i=1}^{k}\Gamma(n_i)}{\Gamma(n+1)} \tag{4.15}$$

As $\pi(\omega_1) = \frac{\Gamma(n)}{\Gamma(n+1)}$ and $\pi(\omega_k) = \frac{\prod_{i=1}^{k}\Gamma(n_i)}{\Gamma(n+1)}$, then:

$$\frac{\pi(\omega_k)}{\pi(\omega_1)} = \frac{\prod_{i=1}^{k}\Gamma(n_i)}{\Gamma(n)} \tag{4.16}$$

When H_0 is true, the Bayes factor will be bigger than 1, but in order to have a standard measure to decide whether combine groups, we propose the use of a transformation that remains in the domain $[0, 1]$ and can be equivalent to $P(H_0)$. Following Casella and Fuentes (2009), we have:

$$P(H_0) = \frac{1}{1+BF}$$

And finally, we will reject the null hypothesis when $P(H_0)$ is smaller than a critical value, typically 0.05 or 0.01. In this case we will separate the partitions, being merged otherwise.

4.2.4 Examples of Bayes factor application

As an example of the performance of the Bayes factor, we will apply it to arbitrary partitions under the existence of one and two groups respectively.

Example 1. Two independent samples:

Two independent samples of sizes $n_1 = n_2 = 100$, are generated from a bivariate normal distributions with means $\mu_1 = (-1, -1)$; $\mu_2 = (1, 1)$ and with covariance matrices $\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/4 \end{bmatrix}$. We arbitrarily split the sample according to the line $X = -0.5 - 0.75X$, as shown in Figure 4.1, to separate the two samples, so we can check if the Bayes factor is able to keep them separate.

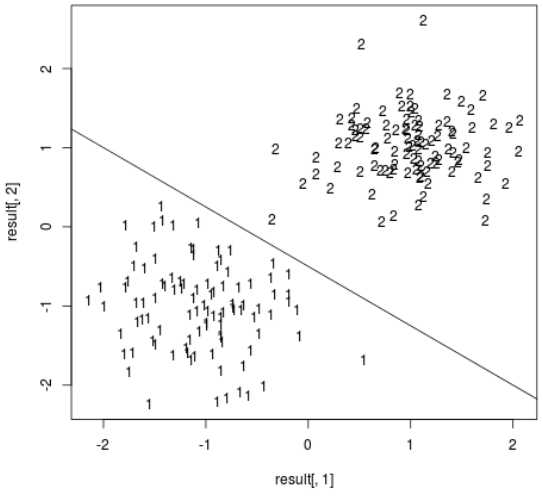

FIGURE 4.1: Bayes factor Example 1, two normal samples

When comparing $H_0 : \kappa = 1$ vs. $H_1 : \kappa = 2$, the following results are obtained:

$$m(X|\omega_2) = 2.662409e - 120, \; m(X|\omega_1) = 2.944844e - 224$$

$$\pi(\omega_1) = 0.005,\ \pi(\omega_2) = 1.115536e-63$$

then:

$$\frac{m(X|\omega_2)}{m(X|\omega_1)} = 9.040915e+103, \frac{\pi(\omega_2)}{\pi(\omega_1)} = 2.231071e-61$$

and finally,

$$BF = 2.017093e+43 \text{ and } p(H_0) = \frac{1}{1+BF} = 4.95763e-44$$

As expected, there is a strong evidence against H_0 and the two groups should be separated.

Example 2: One sample

In this example only one sample of size $n = 200$ is generated from a Normal distribution with mean $\mu_1 = (1, 1)$ and with covariance matrix $\Sigma_1 = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/4 \end{bmatrix}$, while the arbitrary partition on this occasion will take place in the line $X = X$. This configuration is shown in Figure 4.2.

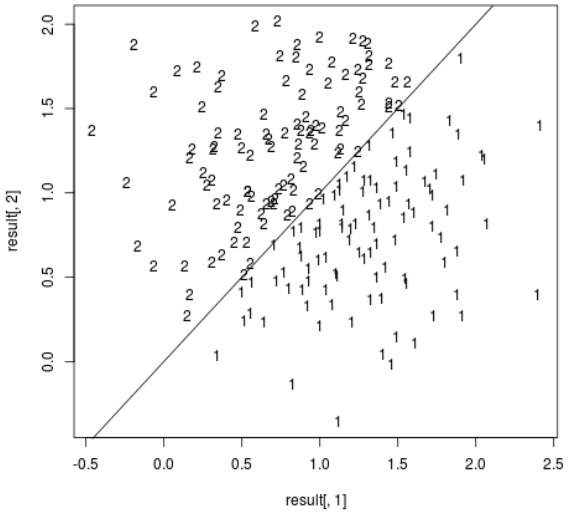

FIGURE 4.2: Bayes factor Example 2, one normal sample

Testing in this occasion $H_0 : \kappa = 1$ vs. $H_1 : \kappa = 2$, we obtain the following results:

$m(X|\omega_2) = 9.515897e-83$, $m(X|\omega_1) = 1.498073e-127$

$\pi(\omega_1) = 0.005$, $\pi(\omega_2) = 1.149685e-63$

where: $\frac{m(X|\omega_2)}{m(X|\omega_1)} = 6.352093e+44$, $\frac{\pi(\omega_2)}{\pi(\omega_1)} = 2.2993e-61$

Finally,

$$BF = 1.460581e-16 \text{ and } P(H_0) = \frac{1}{1+BF} \approx 1$$

Now we have evidence in favour of H_0 so the two partitions should be merged, again as expected. More examples of this test under the framework of the proposed cluster algorithm will be given later in Section 4.4.

4.3 The splitting and group recombining algorithm (SAGRA)

Our algorithm proposal is based on the splitting, cleaning, and recombining processes described in the previous section, and it has as a main goal to return a vector with cluster classes given a multivariate data set. Such classes should help the researcher to unveil the structure of the data, being in this way a tool for exploratory research.

The algorithm holds the usual assumptions for cluster analysis, i.e. every observation is assigned to a group, all observations are classified, and the internal variability of the classes is smaller than between groups. One of the advantages of this procedure is that the data set does not need to be standardized since we use Mahalanobis distances.

The procedure to form the final data configuration is organized in six steps: two splitting steps, one outlier cleaning process and three recombining steps. Each step is detailed in the following subsections.

4.3.1 Split step 1

The input of the algorithm is a sample $x = x_1, x_2, ...x_n$ coming from a p-variate unknown distribution with sample mean $\bar{x}$, and sample covariance matrix S. To start the procedure, the discriminator of each observation is obtained. Recalling Subsection 1.4.2, the discriminator of one data point is the most discrepant point respect to the rest of the sample (See Equation (1.4)).

We also saw in Subsection 1.4.2 that in an univariate sample, the discriminators are the extreme values of the sample, while in multivariate dimensions they lay into the convex hull. When the discriminator function is applied to the data sample, typically all observations will be assigned to a subset of the data points belonging to the convex hull, defining with this step the first split.

To illustrate step by step the behaviour of the proposed clustering procedure, we will apply it to the well-known "Old Faithful" data set introduced in Chapter 1. It considers the waiting time between eruptions and the duration of them from the geyser "Old Faithful" in Yellowstone Park, Wyoming, USA. (Figure 1.1). With the discriminator function, the dataset is split into 8 groups with the distribution given by table 4.1.

TABLE 4.1: Discriminator function distribution for the geyser data set

group	1	2	3	4	5	6	7	8	Total
discriminator observation	19	58	76	149	158	161	197	265	
group size	3	49	79	34	48	4	9	46	272

A minimum size m_0 of the groups is needed to avoid over splitting into small highly homogeneous groups, which can be difficult to merge in the recombining stage of the algorithm. For this reason we set a minimum size equivalent to the 5% of the size of the sample, although it can depend on the complexity of the data set, other alternative could be to use the minimum size proposed by Peña et al. (2004), where $m_0 = p + log(n - p)$. In our example, $n = 272$, so $m_0 = 13.6$.

When this first partition leads to groups such that all of them are of sizes smaller than minimum size, the splitting stop. Otherwise we eliminate the small groups, classifying their observations as isolated observations.

Also the discriminators are extracted from the obtained groups, and temporarily assigned as isolated observations. This is because in deeper levels we want to discover new structures and not define the same partitions in the following steps, as it will happen if we keep the discriminator in. Therefore, the groups which are of sizes smaller than the minimum size are also considered isolated observations, as in the case of groups number 1, 6 and 7, whose sizes are 3, 4, and 9 observations respectively.

As a result of this procedure we obtain a first cluster structure and a set of isolated data points. In the example we get 5 groups with the distribution given by table 4.2.

TABLE 4.2: First splitting step of SAGRA cluster distribution of the geyser example

group	1	2	3	4	5	isolated	Total
size	79	34	46	47	45	21	272

Formally, the step 1 is expressed as:

Require: data set $D = y_1, y_2, ..., y_n$

Split step 1.

for $i = 1 \rightarrow n$ **do**

$$y_l(y_i) \leftarrow arg \max_j \ (y_i - \bar{y}_{(ij)})' \hat{V}^{-1}_{(ij)} (y_i - \bar{y}_{(ij)})$$

end for

$L \leftarrow \{y_j \in D \mid \exists y_i \in D, y_j = y_l(y_i)\}$

Compute $C_1, C_2, ...C_K$, where $K = |L|, i = 1, 2, ..., n$

s.t. $y_i, y_j \in C_k \Leftrightarrow y_l(y_i) = \ y_l(y_j) \forall i, j = 1, 2, ..., n; k = 1, 2, ..., K$

s.t. $|C_k| \geq m_0 \forall k = 1, 2, ..., K$

for $k = 1 \rightarrow K$ **do**

$C_k \leftarrow C_k \setminus L$

end for

Output: $C = \{C_1, C_2, ..., C_K\}$

4.3.2 Split step 2

The second step is to apply the same previous discrimination procedure to each of the previous groups, finding its internal cluster structure.

For each of this "second level" cluster structures, we test if the groups should be split into the basic groups obtained by the splitting, or maintained as in the previous level. We use the recombining test introduced in the previous section setting $H_0 : \kappa = 1$ vs. $H_1 : \kappa = K_i$, being k_i the number of partitions found (second level) in the group i (first level). When $p(H_0) < \alpha$ we reject H_0 and we split into the groups defined by the discrimination of the second level.

All groups which are not split (i.e. $p(H_0) > \alpha$) are separated from the procedure and saved as "candidate groups". These candidate groups will be not split again, and only can be recombined, so they are separated from the rest of the groups. For each of the remaining groups we repeat the procedure until no further partition can be done so we added all sub partitions to the candidate groups.

In the geyser example, the second level partition is shown in table 4.3:

TABLE 4.3: Second splitting step of SAGRA cluster distribution of the geyser example

Level 1	1	1	1	1	2	3	4	5
Level 2	1	2	3	4	1	1	1	1
size	21	20	18	15	33	45	19	21
P-value			1		0.5	0.5	0.5	0.5

The second step splits the group 1 into four subgroups. Then the Bayes factor for this four groups obtains a p-value of 1, indicating that the likelihood of these

partitions is too low, so we do not split this group. For the other groups from the previous split (groups 2 - 5), there is no splitting, so $BF = 1$ and $P(H_0) = 1/2$. Since all p-values in second stage are bigger than $\alpha = 0.01$, we set all those partitions as candidate groups and the splitting procedure is finished and the groups in this step remains as they were in the previous splitting step.

Formally, step 2 is expressed as:

Require: data set $D = y_1, y_2, ..., y_n$, $C = \{C_1, C_2, ..., C_K\}$
  Split step 2.
  $GC \leftarrow \emptyset$
  if $K = 1$ **then**
    $GC \leftarrow D$
  else
    repeat
      $C \leftarrow \emptyset$
      for $i = 1 \rightarrow K$ **do**
        Apply step 1 to C_i' to obtain $C_{1i}', C_{2i}', ..., C_{Ki}'$
        Compute $p(H_0), H_0 : \kappa = 1 \; vs \; H_1 : \kappa = Ki$
        if $p(H_0) > \alpha$ **then**
          $GC \leftarrow GC \cup \{C_i'\}$
        else
          $C \leftarrow C \cup \{C_{1i}'\} \cup \{C_{2i}'\}, ..., \cup \{C_{Ki}'\}$
        end if
      end for
    until $C = \emptyset$
  end if
  Output: $GC = \{GC_1, GC_2, ..., GC_{K'}\}$

4.3.3 Cleaning process

As a result of the two split steps we get a set of "candidate groups" and some isolated data points from the last step. Nevertheless, since a minimum size was established, is possible that some of the candidate groups are still formed by a mix of observations from different clusters. To avoid undesirable recombination due to this misclassified observations, is necessary to apply an "outlier" detection and cleaning process before the recombination steps. The idea is to have pure basic groups with no elements from different clusters.

From the three outliers detection methods we considered in Section 4.2.2, currently our algorithm incorporate the MCD method for efficiency reasons, but future versions of the code will allow the user to choose among those outliers methods.

Coming back to the example, the cleaning is performed inside each group using the MCD method, leading to 16 observations removed from candidate groups. The new group distribution is given by table 4.4:

TABLE 4.4: Cleaning step of SAGRA cluster distribution of the geyser example

group	1	2	3	4	5	isolated	Total
size	77	34	46	40	38	37	272

Formally, the cleaning step is expressed as:

Require: data set $D = y_1, y_2, ..., y_n$, $GC = \{GC_1, GC_2, ..., GC_{K'}\}$

Cleaning step

for $i = 1 \rightarrow K'$ **do**

for $j = 1 \rightarrow n_i$ **do**

Compute RD_j

end for

Compute $c_n(\delta)$

$GC_i \leftarrow GC_i \setminus \{y_{ji} \in GC_i \mid RD_j > c_n(\delta), j = 1, 2, ..n_i\}$

end for

Output: $GC = \{GC_1, GC_2, ..., GC_{K'}\}$

4.3.4 Recombine step 1

The first step in the recombination stage is to order the K' groups such that the bigger partition is labelled as group 1, and the rest depending on how close (using the Mahalanobis distance) they are to the group 1, being "2" the closer, "3" the next, and so on. After ordering, we test for merging groups 1 and 2:

If $p(H_0) \leq \alpha$, we keep them as separated groups, and group 1 will stay as "candidate group". Now we test for merging groups 2 and 3, and so on. If $p(H_0) > \alpha$ We do not split the groups, we relabel the resulting merged group as group 1, and the remaining from 2 to K'-1

The process finishes when just one group remains, and in this case it is also assigned as a new candidate group.

In the example, the groups are merged as shown in table 4.5:

TABLE 4.5: Test results of the first recombining step of SAGRA cluster to the geyser example ($\alpha = 0.01$)

test	1-2	12-3	123-4	4-5
p-value	0.038	1	0	0.003

So three groups are formed with the formers 1-2-3, 4 and 5, with size distribution given by table 4.6.

TABLE 4.6: First recombining step of SAGRA cluster distribution of the geyser example

group	1	2	3	isolated	Total
size	157	40	38	37	272

Formally, the recombine step 1 is expressed as:

Require: data set $D = y_1, y_2, ..., y_n$, $GC = \{GC_1, GC_2, ..., GC_{K'}\}$

Recombine step 1.

$C \leftarrow \emptyset$

$C_1 \leftarrow \arg\max_{GC_k} |GC_k|, k \in 1, 2, ..., K'$

define $C_2, C_3, ...C'_K \mid D_M(C_2, C_1) > D_M(C_3, C_1) > ... > D_M(C'_K, C_1)$

$max \leftarrow K'$

for $k = 1 \rightarrow K' - 1$ **do**

$G \leftarrow C_k \cup C_{k+1}$

Compute in G $p(H_0), H_0 : \kappa = 1 \; vs \; H_1 : \kappa = 2$

if $p(H_0) > \alpha$ **then**

$C_k \leftarrow C_k \cup C_{k+1}$

if $k < max - 1$ **then**

$C_{k+1} \leftarrow C_{k+2}; C_{k+2} \leftarrow C_{k+3}; ...; C_{max-1} \leftarrow C_{max}$

$C_{max} \leftarrow \emptyset$

$max \leftarrow max - 1$

end if

end if

end for

Output: $C = \{C_1, C_2, ..., C_{K''}\}$

4.3.5 Recombine step 2

After we recombine the groups obtained by the splitting process, is still necessary to assign the isolated observations (i.e. discriminators and groups under the minimum size) to one of the candidate groups. This is done simply calculating the Mahalanobis distances (D_M) from each isolated point to all candidates and assigning it to the closer one.

In geyser data, the isolated points where assigned to the three groups as shown in 4.7:

TABLE 4.7: Second recombining step of SAGRA cluster distribution of the geyser example

group	1	2	3	Total
size	176	51	45	272

Formally we have:

Require: data set $D = y_1, y_2, ..., y_n$, $C = \{C_1, C_2, ..., C_{K''}\}$

Recombine step 2.

$D_{isolated} \leftarrow D \setminus C_1 \setminus C_2 ... \setminus C_{K''}$

for $i = 1 \rightarrow |D_{isolated}|$ **do**

$C_j \leftarrow C_j \cup y_i \mid C_j \leftarrow \arg\max_{C_k} D_M(C_k, y_i)$,
$j \in 1, 2, ..., K'', y_i \in D_{isolated}$

end for

Output: $C = \{C_1, C_2, ..., C_{K''}\}$

4.3.6 Recombine step 3

Finally, given that the incorporation of the isolated points increases the variability of the groups, a new merging process (first recombining step) is performed between the candidate groups leading to the final data configuration. The test results for our example is given by table 4.8, while the final data distribution in the two final groups is shown in table 4.9

TABLE 4.8: Test results of the third recombining step of SAGRA cluster for the geyser example

test	1-2	2-3
p-value	0	0.98

The graphical result of the SAGRA procedure applied to the geyser data is shown in the Figure 4.3(a) where we can observe that the two clusters are correctly

TABLE 4.9: Final SAGRA cluster distribution of the geyser example

group	1	2	Total
size	176	96	272

separated, although no error ratios can be calculated because even it is clear that at least two groups are identified in the sample, there are no original labels.

Require: data set $D = y_1, y_2, ..., y_n$, $C = \{C_1, C_2, ..., C_{K''}\}$

Recombine step 3.

Apply Recombine step 1 to C to obtain Final Clusters.

Output: $FC = \{FC_1, FC_2, ..., FC_{K'''}\}$

4.3.7 Comparison with other algorithms

Because the SAGRA continues the work developed by Peña et al. (2004) is natural to compare our results with those obtained by the original SAR algorithm, which results are shown in Figure 4.3(b). Additionally, we will include in the comparison two benchmarking algorithms like k-means (MacQueen 1967), presented in Figure 4.3(c) and M-clust (Fraley and Raftery 1998), plotted in Figure 4.3(d). In the case of k-means we will set the number of groups as the original, two in the case of geyser data.

Regarding those algorithms, SAGRA shows similar results to k-means in detecting two groups, whereas SAR detects also the same two main groups and a group of isolated points between them, while M-clust split one of the main groups into two, leading to a three groups configuration.

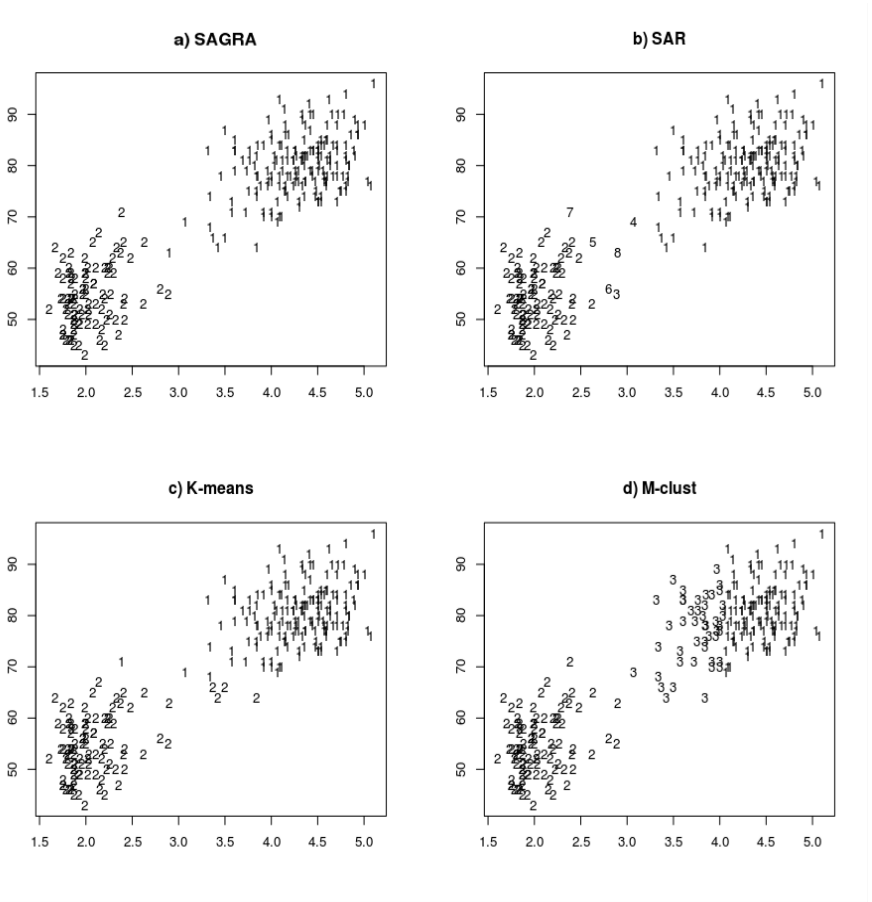

FIGURE 4.3: Comparison of cluster methods applied to the Old Faithful data set

4.3.8 Example: Four independent samples:

In this second example, four independent samples are generated with sizes $n_1 = n_2 = n_3 = n_4 = 100$, coming from a bivariate normal distribution with means $\mu_1 = (2, 2)$, $\mu_2 = (2, -2)$, $\mu_3 = (-2, -2)$, $\mu_4 = (-2, 2)$; and covariance matrices $\Sigma_1 = \begin{bmatrix} 0.25 & 0.15 \\ 0.15 & 0.25 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0.25 & -0.15 \\ -0.15 & 0.25 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 0.25 & 0.15 \\ 0.15 & 0.25 \end{bmatrix}$, $\Sigma_4 = \begin{bmatrix} 0.25 & -0.15 \\ -0.15 & 0.25 \end{bmatrix}$.

In this way we have four well separated groups with different orientations as shown in Figure 4.4, whereas the graphical results of the clustering are in Figure 4.5. In this case, the modified SAR shows similar results to mclust, and slightly better than original SAR, which correctly classify the four groups but creates a 5th small group. K-means split one group into two, and forced by the number of groups set to 4, classify two groups as one.

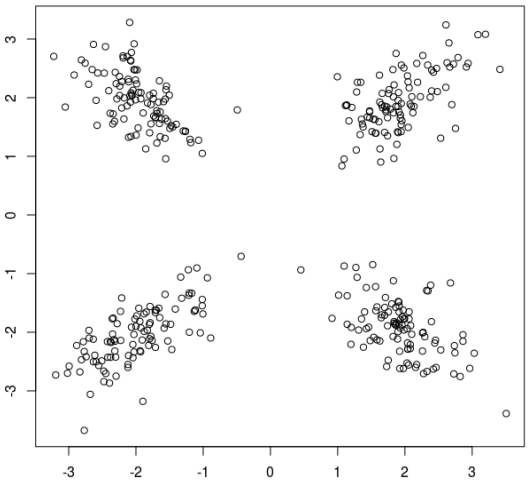

FIGURE 4.4: Four well separated normal samples example

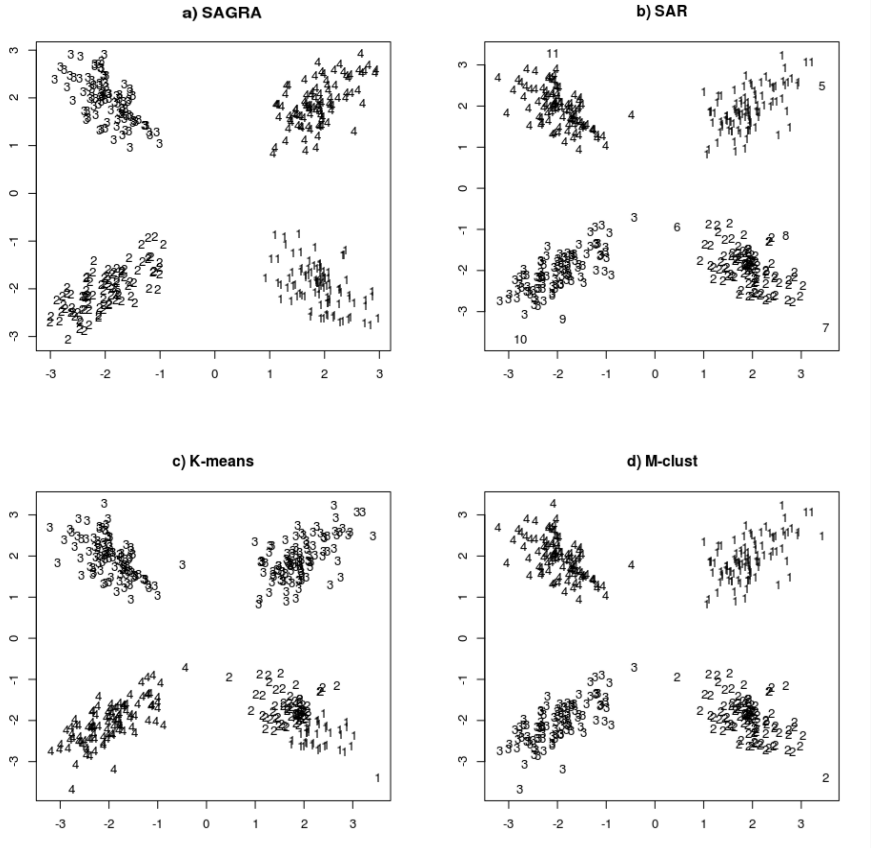

FIGURE 4.5: Comparison of cluster methods applied to four normal simulated samples

4.4 Results

In order to generalize the previous examples and compare the results of the SAGRA algorithm with the other clustering procedures, we use classical measures to evaluate the quality of the output from such class of methods, based on the number of positive and negative decisions when classifying each data point from a data set.

To do so, we simulate data sets to have the original labels, allowing to compare the results from the clustering methods. Notice that generally in cluster analysis the labels are not available, so this comparison can be done only between two different clustering solutions.

Given the total number of pairs of observations $n(n-1)/2$ that can be formed from an original sample of size n, a true positive decision (TP) assigns two observations from the same class in the same cluster, and a true negative (TN) decision assigns two observations from different classes to different clusters.

The errors of the algorithm can be defined in the same way, being a false positive decision (FP) if the algorithm assigns two data points from different classes in the same cluster, while a false negative (FN) decision assigns two observations from the same class in different clusters. The total counting of these four decisions are usually presented in a "Table of Confusion" as given by table 4.1.

Table 4.10: Generic Table of Confusion

	Same cluster	Different clusters
Same class	TP	FN
Different classes	FP	TN

In base of those decisions, we compare the proposed SAGRA method with SAR, K-means and M-clust, using the following measures: Purity, Number of Groups, Rand Index, Adjusted Rand Index, and F_1.

Purity is the sum of the majority of observations assigned to each cluster by the algorithm over the total number of observations, and is a measure of how "pure" the clusters are, in the sense that they are formed only for elements from the

same class. This measure will favour those solutions with many groups but whose elements belongs to the same class, for that reason we include the Number of Groups, so we can observe how close or far the cluster methods are from the original data.

Rand (1971) proposes an index to compute the percentage of correct decisions made by a cluster algorithm, defined by:

$$RI = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.17}$$

A modification of the Rand Index (RI), called Adjusted Rand Index (ARI) was proposed by Hubert and Arabie (1985) to solve some issues from the RI, which expected value is not zero when two random partitions are compared, or that is higher when the number of groups increases. Then, the ARI take values from -1 to 1 and is expressed as:

$$ARI = \frac{\frac{n(n-1)}{2}(TP + FN) - [(TP + FN)(TP + FP) + (FP + TN)(FN + TN)]}{\left[\frac{n(n-1)}{2}\right]^2 - [(TP + FN)(TP + FP) + (FP + TN)(FN + TN)]} \tag{4.18}$$

The F_1 measure is a way to compare the precision and recall when comparing cluster results. Precision is the ratio between the TP over all pairs we assign to the same cluster $P = TP/(TP + FP)$, while Recall is the probability of assigning two elements to the same group given that they are from the same class, $R = TP/(TP + FN)$. Finally, the F_1 is defined as:

$$F_1 = \frac{2 * P * R}{P + R} \tag{4.19}$$

For all of this measures, the more similar the cluster result is respect to the real configuration, the bigger the index will be. The exception of this rule is the "Groups" measure, when the closer to the original number of groups is the best.

The SAGRA algorithm was coded and run under R framework. The code will be soon published in the authors web site, and is also available upon request.

The SAR algorithm was run via the `sarpt` function in Matlab described in Rodriguez (2002), the Mclust algorithm has been run with the R function `Mclust` with models "EII", "VII", "EEI", "VEI", "EVI", "VVI", "EEE", "EEV", "VEV", and "VVV" as a covariance structure, and the possible number of clusters is set to be between 1 and 8. The final configuration is selected by the BIC, as is detailed in Fraley and Raftery (1999). Finally, the K-means algorithm was also run under the R framework, using the function `cascadeKM`, from the `vegan` package where the rule to select the "K" number of clusters in the algorithm is the maximum of the Calinski criteria for $k = 1, ..., 8$ (see Calinski and Harabasz 1974, and Section 1.3).

We test the procedures under four data configurations:

Case 1: Normal distribution

For this case we generate 500 random data sets, each consisting of four independent samples with sizes $n_1 = n_2 = n_3 = n_4 = 100$, coming from a bivariate normal distribution with means $\mu_1 = (1, 1)$, $\mu_2 = (1, -1)$, $\mu_3 = (-1, -1)$, $\mu_4 = (-1, 1)$; and covariance matrices $\Sigma_1 = \begin{bmatrix} 0.25 & 0.15 \\ 0.15 & 0.25 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0.25 & -0.15 \\ -0.15 & 0.25 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 0.25 & 0.15 \\ 0.15 & 0.25 \end{bmatrix}$, $\Sigma_4 = \begin{bmatrix} 0.25 & -0.15 \\ -0.15 & 0.25 \end{bmatrix}$ respectively.

Case 2: Two correlated uniform samples

In a second scenario we generate 500 random data sets, each consisting of two independent samples with sizes $n_1 = n_2 = 500$, coming from a bivariate uniform distribution with means $\mu_1 = (0, 0)$, $\mu_2 = (-0.5, 0)$. The correlation between the two variables on each sample is set to be $\rho = 0.9$.

Case 3: Three geometric uncorrelated uniform samples

Now we generate 500 random data sets, each consisting of three independent samples, with sizes $n_1 = n_2 = n_3 = 500$ on geometric shapes formed by uncorrelated uniform observations in the shape of one circle and two rectangles.

Case 4: Two half moons

Finally for the last case we generate 500 data sets, each consisting of two half moons, with sizes $n_1 = n_2 = 500$, from the R package `spa` (Culp 2011). The two moons are oriented opposite to each other, so they cannot be linearly separated.

One of the samples of each dataset is shown in the Figure 4.6, and the results of the simulations are shown in Table 4.2.

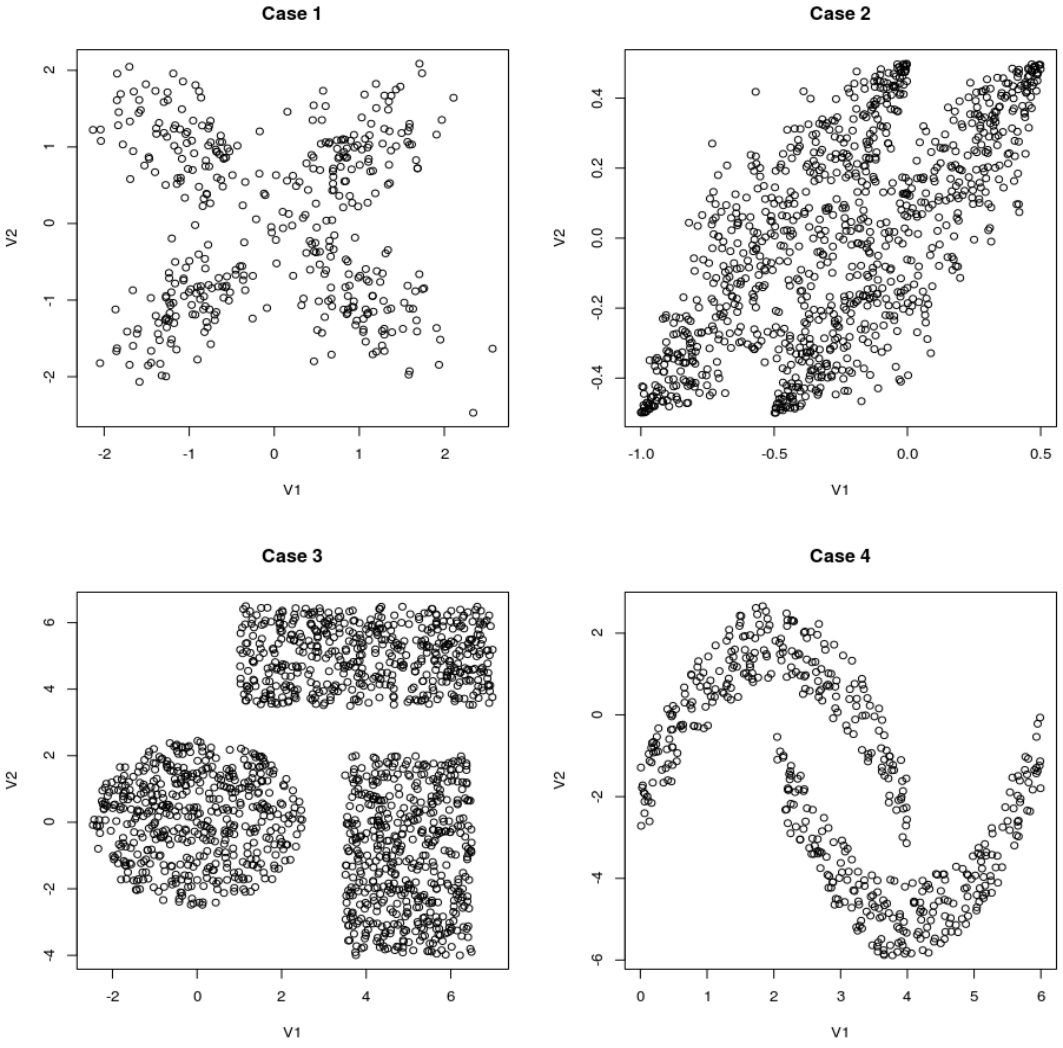

FIGURE 4.6: Four data configurations to test the performance of the SAGRA algorithm

TABLE 4.11: Average quality measures from 500 simulations for each case

	Criteria	SAGRA	SAR	K-means	M-clust
Four Normals	Purity	**0.96**	0.25	0.92	**0.96**
	Groups	4.02	1.37	5.83	**4.00**
	RI	**0.96**	0.25	0.90	**0.96**
	ARI	**0.89**	0	0.70	**0.89**
	F1	**0.92**	0.4	0.76	**0.92**
Two correlated uniforms	Purity	**0.96**	0.5	**0.96**	0.94
	Groups	5.08	**1.11**	7.97	8.13
	RI	**0.72**	0.5	0.61	0.63
	ARI	**0.44**	0	0.21	0.26
	F1	**0.62**	0.67	0.37	0.45
Three geometric uniforms	Purity	**1.00**	0.84	**1.00**	**1.00**
	Groups	3.86	**2.52**	5.39	7.55
	RI	**0.96**	0.89	0.88	0.81
	ARI	**0.90**	0.79	0.69	0.51
	F1	**0.93**	0.88	0.76	0.61
Two half moons	Purity	0.99	0.51	0.94	**1.00**
	Groups	4.76	**1.15**	4.10	6.43
	RI	0.75	0.51	**0.78**	0.67
	ARI	0.51	0.02	**0.56**	0.33
	F1	0.67	0.67	**0.71**	0.50

4.5 Conclusions

The SAGRA (Split And Group Recombining Algorithm) method was developed based on a splitting and recombining methodologies in a similar way as the SAR algorithm proposed by Peña et al. (2004). In comparison with SAR itself and other classical cluster analysis procedures such as K-means and M-clust, SAGRA shows competitive results. We applied those methods over four different data configurations and we took five performance measures (Purity, Groups, RI, ARI, and F1).

SAGRA obtained similar results to M-clust under normally distributed data set (where M-clust tend to be optimal), and in general, better results than other methodologies in the rest of the cases over all the measures, with the exception of the detection of the real number of groups, where no methodology was exact for these four simulated data sets.

Some issues of the algorithm in comparison to other methods need to be considered. First, is necessary to fix two parameters, the minimum size for the splitting process and the critical value for the $p(H_0)$ in the recombining step. The minimum size determines one of the stopping rules in the algorithm, and is required to split the sample in such a way that 1) the groups are small enough to adequately separate the different classes and 2) the groups are big enough so they can be tested to be combined using a model approach using the Bayes factor. As reported in the description of the algorithm, an empirical use of a minimum size equivalent to the 5% of the total sample size adequately holds for these two conditions in a general context, although the number can depend on the complexity of the data set.

Regarding the critical value for the $p(H_0)$, is important to notice that we are comparing two different models, one where the data comes from the same distribution, and other when the data is generated by the structure implied in the partition. We choose the null hypothesis to be that where the data set is coming from the same distribution, so it forms only one group, and we will split the sample only if there are strong evidence for that, choosing a value $p(H_0) < 0.01$ to split the data in the given partitions for our examples.

As advantages, the algorithm does not need to fix the number of groups, as k-means, or it is not necessary to compare different solutions as M-clust or the original SAR algorithm, both of them using the BIC criteria to choose the final data configuration. This is important since the BIC criteria is optimum to compare Normal distributions but tends to overestimate the real number of groups when the data depart from normality.

In summary, although the four data configuration used to test the algorithms are quite different, the obtained results show that the SAGRA algorithm is competitive with respect to the benchmarking algorithms in Cluster Analysis obtaining more than 95% of purity in each example.

Chapter 5

Conclusions and further research

Throughout this thesis we have reviewed and proposed several methods to achieve its original goal: to unveil the hidden structure of a given data set using splitting and recombining approaches.

In Chapter 1 we settled the main concepts we used later. Starting with a brief review of Cluster Analysis methods we revisited the SAR algorithm by Peña et al. (2004) who proposed a methodology to split and recombine a data set to discover clusters. This algorithm served as motivation for our thesis proposal, given that it obtains a set of homogeneous groups that cannot be recombined by classical methods, as a result of the splitting process.

A natural way to address the problem of recombining non-independent data partitions as those obtained by the SAR, is to study the properties of order statistics. In an univariate framework, any partition of the sample into non overlapping groups implies the definition of a certain order, then we started Chapter 2 reviewing some of the main results about order statistics properties such as their distribution and moments. We showed that except for some special cases, there are not closed expressions for such class of statistics. We focus later in the case of the normal distributions, where some approximations have been proposed in the literature, and we proposed the use of the triangular distribution as a possible approximation.

Later on the chapter we focused on linear combination of order statistics, presenting a bootstrap estimation of their moments proposed originally by Hutson and Ernst (2000). We claimed that is possible to use bootstrap linear combination of order statistics as a tool to decide whether two partitions should be combined or not, and the results indicated this combination can be performed even if the partitions are in the tales or in the center of the original generating distribution.

Finally, given the limitations of procedures based on order statistics, we used depth functions as an extension of them in higher dimensions. Results showed that partitions with similar depth levels can be recombined to discover the cluster structure of the data.

In Chapter 3, we presented an univariate methodology for exploratory analysis of recombining partitions. It was based on the study of modes in the density of the data, since departing from unimodality can be a sign of the presence of clusters. Hartigan and Hartigan (1985) derived a test to detect unimodality so we developed an algorithm that integrate this test with a partition and recombination processes, using network visualization for the results. We showed that this can be an useful tool to detect heterogeneity in the data, although we also discussed the use of multivariate mode detection tests to avoid projecting multivariate data into one dimension. The results of the application of such test showed that is possible to detect the cluster structure of the data, although more research can be oriented to estimate the proper fine-tuning of some parameters of the test for a given dataset or distribution.

In Chapter 4 we address the recombining problem in multivariate dimensions. To do so, we use a Bayes Factor test to compare two models, one when the original configuration of the data is given by the partitions, against a single distribution with no clusters explaining the data set.

We showed that the proposed Bayes Factor test is able to combine partitions coming from the same distribution and split those from separate clusters. Using this ratio, we built a cluster analysis algorithm, that we called SAGRA (Split, And Group Recombining Algorithm), integrating the discriminator function by Peña

et al. (2004) for the split process, an outlier cleaning process, and finally the Bayes Factor for merging.

We compared the behaviour and results of our procedure against the original SAR and two popular cluster algorithms: K-means and M-clust. The results showed that our method is competitive and obtain similar or better results that alternatives under different data structures.

In summary, the main contributions of this thesis are:

- The exact and approximate expectations of order statistics from triangular distributions. These results can be used to fit the expectations of order statistics from normal distributions.
- A bootstrap approach to recombine univariate partitions based on linear combinations of order statistics.
- A depth based approach to recombine multivariate partitions.
- A pairwise method to recombine partitions by using unimodality tests both in one or multidimensional data, including a graphical tool to visualize the evolution of the recombining.
- A new clustering algorithm (SAGRA, Splitting and Group Recombining Algorithm) based on a splitting and recombining methodology using the discriminator function and outlier cleaning for splitting, and Bayes Factors for recombining.

Based on these contributions, some possible extensions and further research lines are:

- For the bootstrap methodologies, it is possible to simulate precise cut-offs for several sample sizes and number of partitions assuming normality or other distributions.

- To derive parametric distributions of linear combination of order statistics so a formal statistical test can be performed.

- In the case of the depth recombination, is still open the problem of define the best depth measure, by comparing the performance of several measures over different data configurations.

- To develop alternatives for the discriminator function, for example considering the comparison respect to the median (or the deepest point) instead of the mean, which can increase the robustness of the method.

- To make flexible the normality assumption for the partitions in the Bayes Factor of the SAGRA algorithm. This can improve the results in cluster detection when data is not normally distributed, specially in terms of identify the proper number of groups, where the original SAR have better performance.

- Is necessary to keep improving the efficiency of the algorithm in order to be able of using it for "big data", where clustering tools are highly needed and only few methods have proven be efficient over.

- To develop specific tools for mixed data, i.e. considering numerical and categorical variables at the same time. One option is to include the Gower distance (Gower 1971) in the discriminator function, so these two different types of variables can be integrated in one cluster solution.

- In terms of the recombination methods, besides those open lines already mentioned in the previous points, another interesting direction is to develop model based tools in a similar way as the M-clust algorithm (Yeung et al. 2001), or using a Latent Class Analysis approach (Goodman 1974; Lazarsfeld and Henry 1968).

Cluster analysis in general, and splitting and recombining methodologies in particular, are still open problems in modern statistics, where better and faster algorithms are produced everyday. This research contributed to the problem of

recombine partitions, comparing alternatives, proposing new methodologies and pointing out some open lines where to focus future research.

Appendix A

A.1 Proof of Proposition (2.2.1)

Remembering that the k-th moment of an order statistics with i:n is defined as:

$$E(x^k) = \int_{-\infty}^{\infty} x^k f(x_{i:n}) dx \tag{A.1}$$

where $f(x_{i:n})$ is the distribution function of an i:n order statistics given by:

$$f(x_{i:n}) = \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1}[1-F(x)]^{n-i} f(x) \tag{A.2}$$

with $f(x); F(x)$ the distribution and cumulative function of the original variables.

Let $f(x); F(x)$ the functions given by (2.22) and (2.23), then the k-th moment of the (i:n) order statistic coming from a triangular distribution with mean 0 and range θ will be:

$$\begin{aligned} &E(x^k)= \\ &C\int_{-\theta}^{0} x^k \left[\frac{x^2+2\theta x+\theta^2}{2\theta^2}\right]^{i-1} \left[1-\left(\frac{x^2+2\theta x+\theta^2}{2\theta^2}\right)\right]^{n-i} \left(\frac{x+\theta}{\theta^2}\right) dx \\ &+C\int_{0}^{\theta} x^k \left[\frac{\theta^2+2\theta x-x^2}{2\theta^2}\right]^{i-1} \left[1-\left(\frac{\theta^2+2\theta x-x^2}{2\theta^2}\right)\right]^{n-i} \left(\frac{\theta-x}{\theta^2}\right) dx \\ &=p(x)+q(x) \end{aligned} \tag{A.3}$$

where $C = \frac{n!}{(i-1)!(n-i)!}$

Considering the first addition term p(x):

Let:

$$u = F(x) = \frac{x^2 + 2\theta x + \theta^2}{2\theta^2}$$
$$\Rightarrow \frac{du}{dx} = \frac{2x + 2\theta}{2\theta^2} = \frac{x + \theta}{\theta^2}$$
$$\Rightarrow du = \frac{x + \theta}{\theta^2} dx$$

Integration limits:

$$x = 0 \Rightarrow F(x) = \frac{1}{2} = u$$
$$x = -\theta \Rightarrow F(x) = 0 = u$$

Then $x^k = [F^{-1}(u)]^k$, so we have:

$$\frac{x^2 + 2\theta x + \theta^2}{2\theta^2} = u \Rightarrow x^2 + 2\theta x + \theta^2 - 2\theta^2 u = 0$$
$$\Rightarrow x_1 = -\theta - \theta\sqrt{2u}$$
$$x_2 = -\theta + \theta\sqrt{2u}$$

Its necessary to verify which solution holds the conditions $u = 0 \Rightarrow x = -\theta$; and $u = \frac{1}{2} \Rightarrow x = 0$:

$$\begin{matrix} u = 0 \Rightarrow x_1 = -\theta; x_2 = -\theta \\ u = \dfrac{1}{2} \Rightarrow x_1 = -2\theta; x_2 = 0 \end{matrix} \Rightarrow x_2 = -\theta + \sqrt{2u} = F^{-1}(u) \ .$$

Replacing all terms in p(x) we have:

$$p(x) = C \int_0^{1/2} \left[-\theta + \theta\sqrt{2u}\right]^k u^{i-1}[1-u]^{n-i} du$$

In the case k=1, $(E[x_{i:n}])$ is:

$$p(x)=C\int_0^{1/2}\left[-\theta+\theta\sqrt{2u}\right]u^{i-1}[1-u]^{n-i}du$$

$$p(x)=C\int_0^{1/2}\theta\sqrt{2u}u^{i-1}[1-u]^{n-i}-\theta u^{i-1}[1-u]^{n-i}du$$

$$p(x)=C\int_0^{1/2}\theta\sqrt{2u}u^{i-1}[1-u]^{n-i}du-C\int_0^{1/2}\theta u^{i-1}[1-u]^{n-i}du$$

$$p(x)=C\theta\sqrt{2}\int_0^{1/2}u^{i-\frac{1}{2}}[1-u]^{n-i}du-C\theta\int_0^{1/2}u^{i-1}[1-u]^{n-i}du$$

$$p(x)=\sqrt{2}C\theta\cdot B\left(\frac{1}{2};i+\frac{1}{2},n-i+1\right)-C\theta\cdot B\left(\frac{1}{2};i,n-i+1\right) \tag{A.4}$$

Where $B(x,a,b)$ is the incomplete beta function defined as:

$$B(x,a,b)=\int_0^x t^{a-1}(1-t)^{b-1}dt$$

Analogously we obtain that the second term q (x) will be:

$$q(x)=C\theta\cdot B\left(\frac{1}{2};n-i+1,i\right)-C\theta\sqrt{2}\cdot B\left(\frac{1}{2};n-i+\frac{3}{2},i\right) \tag{A.5}$$

so far the expectation is:

$$\begin{gathered}E(x_{i:n})=\\ \sqrt{2}C\theta\cdot B\left(\frac{1}{2};i+\frac{1}{2},n-i+1\right)-C\theta\cdot B\left(\frac{1}{2};i,n-i+1\right)+\\ C\theta\cdot B\left(\frac{1}{2};n-i+1,i\right)-\sqrt{2}C\theta\cdot B\left(\frac{1}{2};n-i+\frac{3}{2},i\right)\end{gathered} \tag{A.6}$$

or, equivalently:

$$\begin{aligned} E(x_{i:n}) = \sqrt{2}C\theta \left[B\left(\frac{1}{2}; i+\frac{1}{2}, n-i+1\right) - B\left(\frac{1}{2}; n-i+\frac{3}{2}, i\right)\right] \\ + C\theta \left[B\left(\frac{1}{2}; n-i+1, i\right) - B\left(\frac{1}{2}; i, n-i+1\right)\right] \end{aligned} \tag{A.7}$$

Remembering that C is the term:

$$C = \frac{n!}{(i-1)!(n-i)!} = B(i, n-i+1)^{-1} = B(n-i+1, i)^{-1} \tag{A.8}$$

The regularized incomplete beta function is defined as the quotient:

$$I_x(a,b) = \frac{B(x;a,b)}{B(a,b)}$$

Therefore, we have:

$$B(x;a,b) = I_x(a,b) \cdot B(a,b) \tag{A.9}$$

where $B(a,b)$ is the standard Beta function.

Using (A.8) and (A.9), Equation (A.6) can be rewritten as:

$$\begin{aligned} E(x_{i:n}) = \sqrt{2}C\theta \left[B\left(\frac{1}{2}; i+\frac{1}{2}, n-i+1\right) - B\left(\frac{1}{2}; n-i+\frac{3}{2}, i\right)\right] \\ + \theta \left[I_{\frac{1}{2}}\left(n-i+1, i\right) - I_{\frac{1}{2}}\left(i, n-i+1\right)\right] \end{aligned} \tag{A.10}$$

Using one of the properties of the regularized incomplete beta function, $I_x(a,b) = 1 - I_{1-x}(b,a)$ we have:

$$E(x_{i:n}) = \sqrt{2}C\theta\left[B\left(\frac{1}{2}; i+\frac{1}{2}, n-i+1\right) - B\left(\frac{1}{2}; n-i+\frac{3}{2}, i\right)\right] + \theta\left[1 - 2I_{\frac{1}{2}}\left(i, n-i+1\right)\right] \tag{A.11}$$

Or, equivalently:

$$E(x_{i:n}) = \sqrt{2}C\theta\left[B\left(\frac{1}{2}; i+\frac{1}{2}, n-i+1\right) - B\left(\frac{1}{2}; n-i+\frac{3}{2}, i\right)\right] + \theta\left[1 - 2\frac{B\left(\frac{1}{2}, i, n-i+1\right)}{B\left(i, n-i+1\right)}\right] \tag{A.12}$$

A.2 Proof of Proposition (2.2.2)

Equation (2.24) has four components expressed as incomplete betas:

a) $B\left(\frac{1}{2}; i+\frac{1}{2}, n-i+1\right)$

b) $B\left(\frac{1}{2}; n-i+\frac{3}{2}, i\right)$

c) $B\left(\frac{1}{2}; n-i+1, i\right)$

d) $B\left(\frac{1}{2}; i, n-i+1\right)$

Where b) and c) have i as third parameter of the incomplete beta, and a) and d) have $n-i+1$. We will show how to approximate these equations for extreme values of i ($i < \frac{n}{2}$).

The plot of beta function associated to (a) for values of $i = 1$, $n = 4$ is shown in figure A.1.

In this case the incomplete beta value, that is, the area under the curve from 0 to 0.5 is equal to 0.0895, while the value of the traditional beta function (area under the entire curve) is equal to 0.1016, with a difference between the two values of 0.0121.

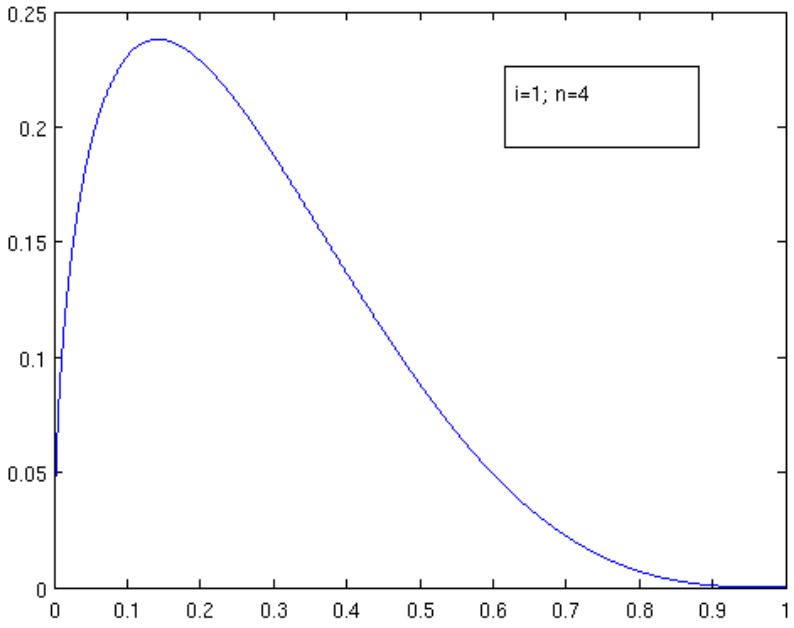

FIGURE A.1: Function $B\left(x; i+\frac{1}{2}, n-i+1\right)$, for i=1, n=4

This difference decreases as n increases, for example, for $n = 10$, the curve is represented in figure A.2. Where the difference between the two (traditional beta and incomplete beta) is close to 0.

In the case of the Beta function associated to (b), and for $i = 1$, $n = 4$ the graph will be given by A.3. In this case the approximation to a traditional beta is not the most suitable, since there is a great difference between the area under the curve to the value 0.5 and the entire area. However, the incomplete beta value (the integral from 0 to 0.5) is quite small, in this case 0.0098.

As n increases, this value is becoming smaller, for example to $n = 10$, we will have the plot given by figure A.4

Where the value of the incomplete Beta function defined in b) (area under the curve from 0 to 0.5) is zero.

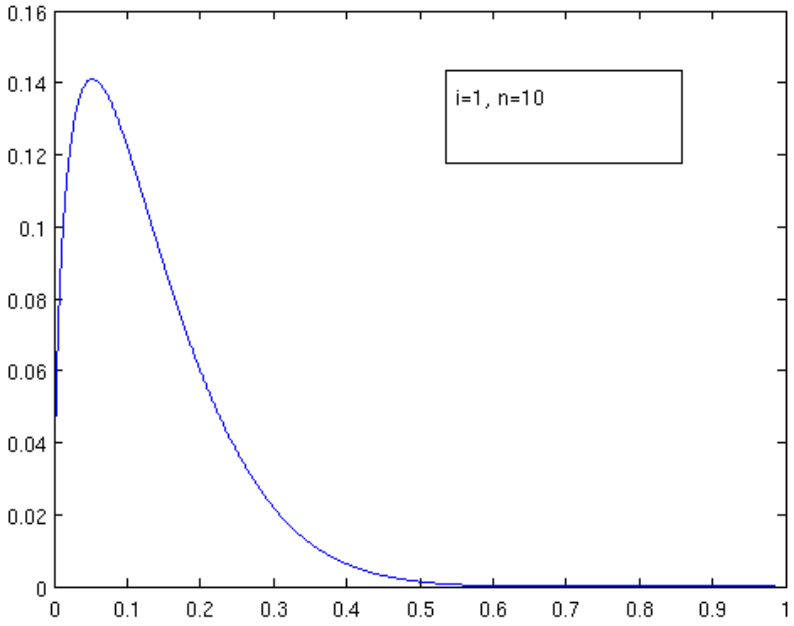

FIGURE A.2: Function $B\left(x; i+\frac{1}{2}, n-i+1\right)$, for i=1, n=4

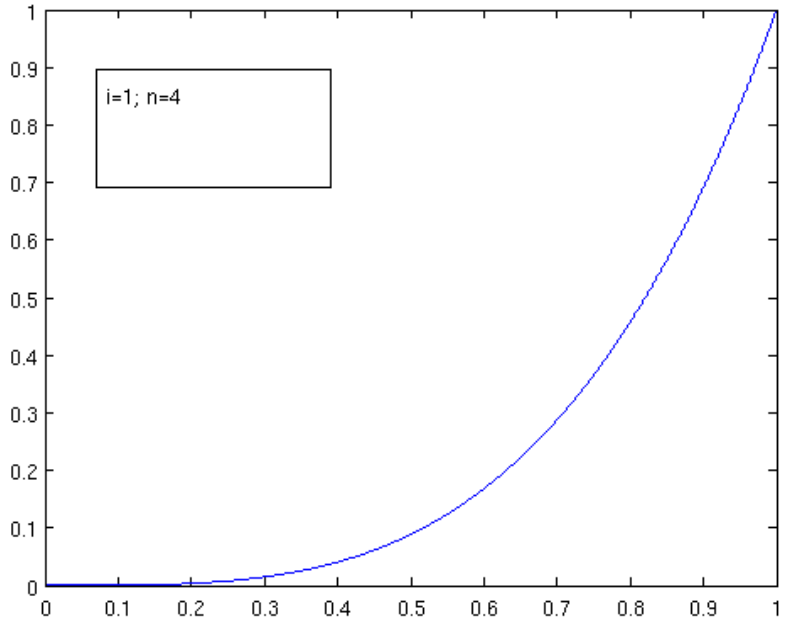

FIGURE A.3: Function $B\left(x; n-i+\frac{3}{2}, i\right)$, for i=1, n=4

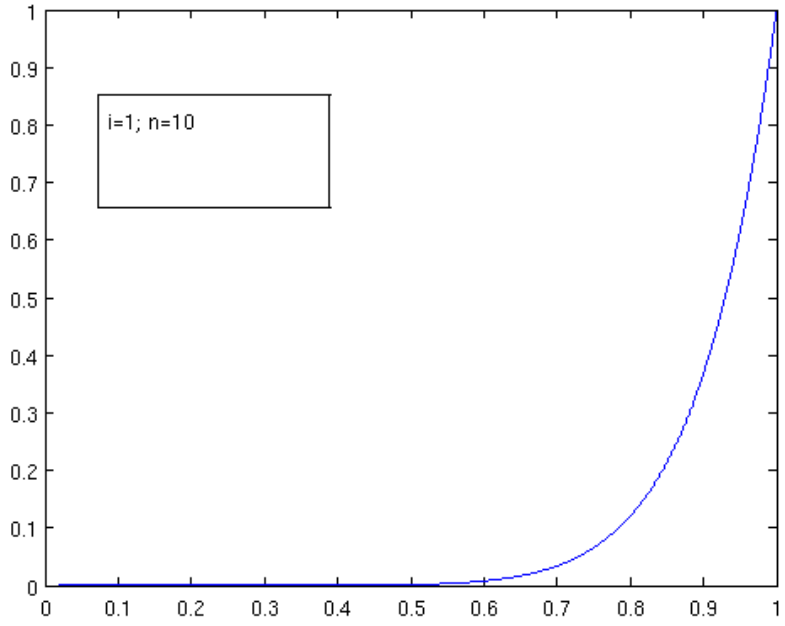

FIGURE A.4: Function $B\left(x; n-i+\frac{3}{2}, i\right)$, for i=1, n=10

So we have that for non-central values of i ($i << \frac{n}{2}$), and with a large enough n (>10),

a) $B\left(\frac{1}{2}; i+\frac{1}{2}, n-i+1\right) \sim B\left(i+\frac{1}{2}, n-i+1\right)$

b) $B\left(\frac{1}{2}; n-i+\frac{3}{2}, i\right) \sim 0$

c) $B\left(\frac{1}{2};, n-i+1, i\right) \sim 0$

d) $B\left(\frac{1}{2}; i, n-i+1\right) \sim B\left(i, n-i+1\right)$

And Equation (A.7) can be approximated by:

$$E_{aprox}(x_{i:n}) = \sqrt{2} C\theta B\left(i+\frac{1}{2}, n-i+1\right) - C\theta B\left(i, n-i+1\right) \tag{A.13}$$

, and because C is the term: $B(i, n-1+1)^{-1}$, we have:

$$E_{aprox}(x_{i:n}) = \frac{\sqrt{2}\theta B\left(i+\frac{1}{2}, n-i+1\right)}{B\left(i, n-i+1\right)} - \theta;\ i << \frac{n}{2} \tag{A.14}$$

Is easy to see in Equation (A.7) that $E(x_{i:n}) = -E(x_{n-i+1:n})$, assessing the symmetry of the triangular distribution. Then for the right part of the distribution when $i >> \frac{n}{2}$ we have:

$$E_{aprox}(x_{i:n}) = \theta - \frac{\sqrt{2}\theta B\left(n-i+\frac{3}{2}, i\right)}{B\left(n-i+1, i\right)};\ i >> \frac{n}{2} \tag{A.15}$$

Leading to Equation (2.25).

Bibliography

Aggarwal, C. C. (2013), *Outlier Analysis*, New York: Springer.

Agostinelli, C. and Romanazzi, M. (2009), "localdepth: Local Depth," *R package version 0.5-4*.

Ahmed, M. O. and Walther, G. (2012), "Investigating the Multimodality of Multivariate Data with Principal Curves," *Computational Statistics & Data Analysis*, 56, 4462–4469.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (2008), *A First Course in Order Statistics*, Philadelphia: SIAM.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (2009a), "Multivariate Order Statistics via Multivariate Concomitants," *Journal of Multivariate Analysis*, 100, 946–951.

— (2009b), "On Multivariate Order Statistics. Application to Ranked Set Sampling," *Computational Statistics & Data Analysis*, 53, 4555–4569.

Atkinson, A. C. and Riani, M. (2007), "Exploratory Tools for Clustering Multivariate Data," *Computational Statistics & Data Analysis*, 52, 272–285.

Azzalini, A. and Bowman, A. W. (1990), "A Look at Some Data on the Old Faithful Geyser," *Applied Statistics*, 39, 357–365.

Bairamov, I. and Gebizlioglu, O. (1997), "On the Ordering of Random Vectors in a Norm Sense," *Journal of Applied Statistical Science*, 6, 77–86.

Balakrishnan, N., Charalambides, C. A., and Papadatos, N. (2003), "Bounds on Expectation of Order Statistics from a Finite Population," *Journal of Statistical Planning and Inference*, 113, 569–588.

Ball, G. H. and Hall, D. J. (1965), "ISODATA, a Novel Method of Data Analysis and Pattern Classification," Tech. Rep. AD699616, Stanford Research Institute.

Banfield, J. D. and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010), "Combining Mixture Components for Clustering," *Journal of Computational and Graphical Statistics*, 9, 332–353.

Bendre, S. M. and Kale, B. K. (1985), "Masking Effect on Tests for Outliers in Exponential Models," *Journal of the American Statistical Association*, 80, 1020–1025.

— (1987), "Masking Effect on Tests for Outliers in Normal Samples," *Biometrika*, 74, 891–896.

Bickel, P. J. and Fan, J. (1996), "Some Problems on the Estimation of Unimodal Densities," *Statistica Sinica*, 6, 23–45.

Biernacki, C., Celeux, G., and Govaert, G. (2000), "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.

Blom, G. (1958), *Statistical Estimates and Transformed Beta-Variables*, New York: John Wiley & Sons.

Branch, J. W., Giannella, C., Szymanski, B., Wolff, R., and Kargupta, H. (2012), "In-Network Outlier Detection in Wireless Sensor Networks," *Knowledge and Information Systems*, 34, 23–54.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Statistics/Probability Series, New York: Chapman and Hall.

Brown, B. (1983), "Statistical Uses of the Spatial Median," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 45, 25–30.

Burman, P. and Polonik, W. (2009), "Multivariate Mode Hunting: Data Analytic Tools with Measures of Significance," *Journal of Multivariate Analysis*, 100, 1198–1218.

Calinski, T. and Harabasz, J. (1974), "A Dendrite Method for Cluster Analysis," *Communications in Statistics Theory and Methods*, 3, 1–27.

Cascos, I., López, A., and Romo, J. (2011), "Data Depth in Multivariate Statistics," *Boletín de Estadística e Investigación Operativa*, 27, 151–174.

Casella, G. and Fuentes, C. (2009), "Testing for the Existence of Clusters," *Statistics and Operations Research Transactions*, 33, 115–146.

Caussinus, H. and Ruiz-Gazen, A. (1994), "Projection Pursuit and Generalized Principal Component Analysis," in *New Directions in Statistical Data Analysis and Robustness*, eds. Morgenthaler, S., Ronchetti, E., and Stahel, W., Basel: Birkhuser Verlag, pp. 35–46.

— (1995), "Metrics for Finding Typical Structures by Means of Principal Component Analysis," in *Data Science and its Applications*, eds. Escoufier, Y. and Hayashi, C., Tokyo: Academy Press, pp. 177–192.

Chau, K. W. (1995), "The Validity of the Triangular Distribution Assumption in Monte Carlo Simulation of Construction Costs: Empirical Evidence from Hong Kong," *Construction Management and Economics*, 13, 15–21.

Chaudhuri, P. (1996), "On a Geometric Notion of Quantiles for Multivariate Data," *Journal of the American Statistical Association*, 91, 862–872.

Cheng, Y. (1995), "Mean Shift, Mode Seeking, and Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17, 790–799.

Cho, H. Y., Oh, J. H., Kim, K. O., and Shim, J. S. (2013), "Outlier Detection and Missing Data Filling Methods for Coastal Water Temperature Data," *Proceedings of the 12th International Coastal Symposium (Plymouth, England), Journal of Coastal Research, Special Issue No. 65*, 1898–1903.

Cuesta-Albertos, J., Gordaliza, A., and Matrán, C. (1997), "Trimmed K-means: An Attempt to Robustify Quantizers," *The Annals of Statistics*, 25, 553–576.

Cuevas, A., Febrero, M., and Fraiman, R. (2000), "Estimating the Number of Clusters," *Canadian Journal of Statistics*, 28, 367–382.

Culp, M. (2011), "spa: Semi-Supervised Semi-Parametric Graph-Based Estimation in R," *Journal of Statistical Software*, 40, 1–29.

David, H. A. and Nagaraja, H. N. (1970), *Order Statistics*, Hoboken (New Jersey): John Wiley & Sons.

Diday, E. (1973), "The Dynamic Clusters Method in Nonhierarchical Clustering," *International Journal of Parallel Programming*, 2, 61–88.

Ding, Y., Dang, X., Peng, H., and Wilkins, D. (2007), "Robust Clustering in High Dimensional Data Using Statistical Depths." *BMC Bioinformatics*, 8.

Donoho, D. and Gasko, M. (1992), "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness," *The Annals of Statistics*, 20, 1803–1827.

Dunn, J. (1973), "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, 3, 32–57.

Dyckerhoff, R. (2004), "Data Depths Satisfying the Projection Property," *Allgemeines Statistisches Archiv*, 88, 163–190.

Einbeck, J. (2011), "Bandwidth Selection for Mean-Shift Based Unsupervised Learning Techniques: a Unified Approach Via Self-Coverage," *Journal of pattern recognition research*, 6, 175–192.

Einbeck, J. and Evers, L. (2012), "LPCM: Local Principal Curve Methods," *R package version 0.44-6*.

Filzmoser, P., Garret, R., and Reimann, C. (2005), "Multivariate Outlier Detection in Exploration Geochemistry," *Computers & geosciences*, 31, 579–587.

Filzmoser, P. and Gschwandtner, M. (2013), "mvoutlier: Multivariate Outlier Detection Based on Robust Methods," *R package version 1.9.9*.

Fraiman, R., Ghattas, B., and Svarc, M. (2011), "Interpretable Clustering Using Unsupervised Binary Trees," *Arxiv preprint arXiv:1103.5339*, 1–22.

Fraiman, R. and Meloche, J. (1999), "Multivariate L-estimation," *Test*, 8, 255–317.

Fraley, C. and Raftery, A. (1998), "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578–588.

Fraley, C. and Raftery, A. E. (1999), "MCLUST: Software for Model-Based Cluster and Discriminant Analysis," *Journal of Classification*, 16, 297–306.

— (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

Friedman, J. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, 82, 249–266.

Friedman, J. and Tukey, J. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers*, c-23, 881 – 890.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, San Diego: Academic Press.

Gan, G., Ma, C., and Wu, J. (2007), *Data Clustering: Theory, Algorithms, and Applications*, Philadelphia: SIAM, Society for Industrial and Applied Mathematics.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008), "A General Trimming Approach to Robust Cluster Analysis," *The Annals of Statistics*, 36, 1324–1345.

— (2010), "A Review of Robust Clustering Methods," *Advances in Data Analysis and Classification*, 4, 89–109.

Geisser, S. (1964), "Posterior Odds for Multivariate Normal Classifications," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 26, 69–76.

Genest, M., Masse, J.-C., and Plante, J.-F. (2012), "depth: Depth Functions Tools for Multivariate Analysis," *R package version 2.0-0*.

Goodman, L. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika*, 61, 215–231.

Gower, J. (1971), "A General Coefficient of Similarity and Some of its Properties," *Biometrics*, 27, 857–871.

Harter, H. (1961), "Expected Values of Normal Order Statistics," *Biometrika*, 48, 151–165.

Hartigan, J. (1975), *Clustering Algorithms (Probability & Mathematical Statistics)*, New York: John Wiley & Sons Inc.

— (1988), "The SPAN Test for Unimodality." in *Classification and Related Methods of Data Analysis*, ed. Book, H. H., Amsterdam: North-Holland Publishing Company, pp. 229 – 236.

Hartigan, J. and Mohanty, S. (1992), "The RUNT Test for Multimodality," *Journal of Classification*, 9, 63–70.

Hartigan, J. J. and Hartigan, P. P. (1985), "The DIP Test of Unimodality," *The Annals of Statistics*, 13, 70–84.

Hartigan, P. (1985), "Algorithm AS 217: Computation of the Dip Statistic to Test for Unimodality," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34, 320–325.

Hennig, C. (2010a), "Methods for Merging Gaussian Mixture Components," *Advances in Data Analysis and Classification*, 4, 3–34.

— (2010b), "Ridgeline Plot and Clusterwise Stability as Tools for Merging Gaussian Mixture Components," in *Classification as a Tool for Research*, eds. Locarek-Junge, H. and Weihs, C., Berlin: Springer, pp. 109–116.

Hodge, V. J. and Austin, J. (2004), "A Survey of Outlier Detection Methodologies," *Artificial Intelligence Review*, 22, 85–126.

Hubert, L. and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218.

Hutson, A. D. and Ernst, M. D. (2000), "The Exact Bootstrap Mean and Variance of an L-Estimator," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 89–94.

Jain, A. and Murty, M. (1999), "Data Clustering: a Review," *ACM computing surveys (CSUR)*, 31, 264 – 323.

Jain, A. K. (2010), "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognition Letters*, 31, 651–666.

Johnson, D. (2002), "The Triangular Distribution as a Proxy for the Beta Distribution in Risk Analysis," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46, 387–398.

Kaluszka, M. and Okolewski, a. (2005), "Bounds for L -Statistics from Weakly Dependent Samples of Random Length," *Communications in Statistics - Theory and Methods*, 34, 1899–1910.

Kaufman, L. and Rousseeuw, P. (1990), *Finding Groups in Data: an Introduction to Cluster Analysis*, Hoboken (New Jersey): John Wiley & Sons.

Koshevoy, G. and Mosler, K. (1997), "Zonoid Trimming for Multivariate Distributions," *The Annals of Statistics*, 25, 1998–2017.

Kothari, V., Wei, I., and Shankar, S. (2013), "Outlier Kinase Expression by RNA Sequencing as Targets for Precision Therapy," *Cancer Discovery*, 3, 280 –293.

Kotz, S. and van Dorp, J. R. (2004), "The Triangular Distribution," in *Beyond Beta*, Singapore: World Scientific Publishing, chap. 1, pp. 1 – 32.

Lazarsfeld, P. and Henry, N. (1968), *Latent structure analysis*, Boston: Houghton Mifflin.

Li, J. (2005), "Clustering Based on a Multilayer Mixture Model," *Journal of Computational and Graphical Statistics*, 14, 547–568.

Liu, R. (1988), "On a Notion of Simplicial Depth," *Proceedings of the National Academy of Sciences*, 85, 1732–1734.

— (1990), "On a Notion of Data Depth Based on Random Simplices," *The Annals of Statistics*, 18, 405–414.

Liu, R. and Singh, K. (1993), "A Quality Index Based on Data Depth and Multivariate Rank Tests," *Journal of the American Statistical Association*, 421, 252–260.

Lloyd, S. (1982), "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, 28, 129–137.

López-Pintado, S. and Romo, J. (2009), "On the Concept of Depth for Functional Data," *Journal of the American Statistical Association*, 104, 718–734.

López-Pintado, S., Romo, J., and Torrente, A. (2010), "Robust Depth-Based Tools for the Analysis of Gene Expression Data." *Biostatistics*, 11, 254–264.

Lopez-Pintado, S. and Torrente, A. (2013), "Package 'depthTools'," *R package version 0.4*.

MacQueen, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. Le Cam, L. M. and Neyman, J., California, USA, University of California Press, pp. 281–297.

Maechler, M. (2013), "diptest: Hartigan's dip Test Statistic for Unimodality - Corrected Code," *R package version 0.75-5*.

Mahalanobis, P. (1936), "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Sciences (Calcutta)*, 2, 49–55.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

McLachlan, G. and Peel, D. (2004), *Finite Mixture Models*, Wiley Series in Probability and Statistics, New York: John Wiley & Sons.

Moll, V. (2012), "The Stirling Numbers of the Second Kind," in *Numbers and functions: from a classical-experimental mathematician point of view*, American Mathematical Society, pp. 191–209.

Murphy, R. (1951), "On Tests for Outlying Observations," Ph.D. thesis, Princeton University.

Oja, H. (1983), "Descriptive Statistics for Multivariate Distributions," *Statistics & Probability Letters*, 1, 327–332.

Pahuja, D. and Yadav, R. (2013), "Outlier Detection for Different Applications: Review," *International Journal of Engineering*, 2.

Parr, W. and Schucany, W. (1982), "Jackknifing L-statistics with Smooth Weight Functions," *Journal of the American Statistical Association*, 77, 629–638.

Peña, D. and Prieto, F. J. (2001), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," *Technometrics*, 43, 286–310.

Peña, D., Prieto, F. J., and Viladomat, J. (2010), "Eigenvectors of a Kurtosis Matrix as Interesting Directions to Reveal Cluster Structure," *Journal of Multivariate Analysis*, 101, 1995–2007.

Peña, D., Rodriguez, J., and Tiao, G. (2004), "A General Partition Cluster Algorithm," in *COMPSTAT: Proceedings in Computational Statistics: 16th Symposium held in Prague, Czech Republic, 2004*, Springer, pp. 371–379.

Peña, D. and Tiao, G. C. (2006), "The SAR Procedure: A Diagnostic Analysis of Heterogeneous Data," Tech. rep., Universidad Carlos III de Madrid.

Peña, D., Viladomat, J., and Zamar, R. H. (2012), "Nearest-Neighbors Medians Clustering," *Statistical Analysis and Data Mining*, 5, 349–362.

Pitman, J. (1996), "Some Developments of the Blackwell-MacQueen Urn Scheme," *Lecture Notes-Monograph Series*, 30, 245–267.

Popović, B., Janev, M., Pekar, D., Jakovljević, N., Gnjatović, M., Sečujski, M., and Delić, V. (2012), "A Novel Split-and-Merge Algorithm for Hierarchical Clustering of Gaussian Mixture Models," *Applied Intelligence*, 37, 377–389.

Raftery, A. and Dean, N. (2006), "Variable Selection for Model-Based Clustering," *Journal of the American Statistical Association*, 101, 168–178.

Rand, W. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846–850.

Rodriguez, J. (2002), "Contribuciones al Estudio de la Heterogeneidad y la Dependencia," Ph.D. thesis, Universidad Carlos III de Madrid.

Rousseeuw, P. J. (1985), "Multivariate Estimation with High Breakdown Point," in *Mathematical Statistics and Applications, Vol. B*, eds. Grossman, W., Pflug, G., Vincze, I., and Wertz, W., Dordrecht: Reidel, pp. 283 – 297.

Rousseeuw, P. J. and Driessen, K. V. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212 – 223.

Royston, J. (1982), "Algorithm AS 177: Expected Normal Order Statistics (Exact and Approximate)," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31, 161–165.

Rozál, G. and Hartigan, J. (1994), "The MAP Test for Multimodality," *Journal of Classification*, 11, 5–36.

Rychlik, T. (2004), "Optimal Bounds on L-Statistics Based on Samples Drawn with Replacement from Finite Populations," *Statistics*, 38, 391–412.

Sarhan, A. and Greenberg, B. (1956), "Estimation of Location and Scale Parameters by Order Statistics From Singly and Doubly Censored Samples," *The Annals of Mathematical Statistics*, 27, 427 – 451.

Sarhan, A. E. and Greenberg, B. G. (eds.) (1962), *Contributions to Order Statistics*, New York - London: John Wiley & Sons.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Serfling, R. (2012), "Depth," in *Encyclopedia of Environmetrics*, eds. El-Shaarawi, A. H. and Piegorsch, W. W., John Wiley & Sons Inc, 2nd ed., pp. 636–641.

Silverman, B. (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 43, 97–99.

Singh, K. (1991), "A Notion of Majority Depth," Tech. rep., Rutgers University, Department of Statistics.

Steinhaus, H. (1956), "Sur la Division des Corp Materiels en Parties," *Bulletin of the Polish Academy of Sciences*, 1, 801–804.

Stigler, S. (1969), "Linear Functions of Order Statistics," *The Annals of Mathematical Statistics*, 40, 770–788.

Tantrum, J., Murua, A., and Stuetzle, W. (2003), "Assessment and Pruning of Hierarchical Model Based Clustering," in *Proceedings of the ninth SIGKDD*, New York, New York, USA: ACM Press, pp. 197–205.

Teichroew, D. (1956), "Tables of Expected Values of Order Statistics and Products of Order Statistics for Samples of Size Twenty and Less From the Normal Distribution," *The Annals of Mathematical Statistics*, 27, 410.

Tibshirani, R., Walther, G., and Hastie, T. (2001), "Estimating the Number of Clusters in a Data Set via the Gap Statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 411–423.

Tietjen, G. L., Kahaner, D. K., and Beckman, R. J. (1976), "Variances and Covariances of the Normal Order Statistics for Sample Sizes 2 to 50," Tech. rep., Los Alamos Scientific Lab., NM. Los Angeles, USA.

Tukey, J. (1975), "Mathematics and the Picturing of Data," in *Proceedings of the International Congress of Mathematicians*, Vancouver.

Tyler, D. E., Critchley, F., Dümbgen, L., and Oja, H. (2009), "Invariant Coordinate Selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 549–592.

Vardi, Y. and Zhang, C. (2000), "The Multivariate L1-Median and Associated Data Depth," *Proceedings of the National Academy of Sciences*, 97, 1423–1426.

Wang, X., Qiu, W., and Zamar, R. H. R. (2007), "CLUES: A non-Parametric Clustering Method Based on Local Shrinking," *Computational Statistics & Data Analysis*, 52, 286–298.

Ward Jr., J. H. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236–244.

Wei, Y. and McNicholas, P. D. (2012), "Mixture Model Averaging for Clustering and Classification," *arXiv preprint arXiv:1212.5760.*

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001), "Model-Based Clustering and Data Transformations for Gene Expression Data." *Bioinformatics*, 17, 977–87.

Zhang, J. (2002), "Some Extensions of Tukey's Depth Function," *Journal of Multivariate Analysis*, 82, 134–165.

Zuo, Y. and Serfling, R. (2000), "General Notions of Statistical Depth Function," *Annals of Statistics*, 28, 461–482.